Automated detection of reflection in texts. A machine learning based approach

Thesis

AUTOMATED DETECTION OF REFLECTION IN TEXTS

A machine learning based approach

THOMAS DANIEL ULLMANN

Knowledge Media Institute

The Open University

April 2015

A thesis submitted for the degree of Doctor of Philosophy

Thomas Daniel Ullmann: *Automated detection of reflection in texts.* A machine learning based approach. April 2015.

SUPERVISORS:

Prof. Dr. Peter Scott

Dr. Fridolin Wild

ABSTRACT

Promoting reflective thinking is an important educational goal. A common educational practice is to provide opportunities for learners to express their reflective thoughts in writing. The analysis of such text with regard to reflection is mainly a manual task that employs the principles of content analysis.

Considering the amount of text produced by online learning systems, tools that automatically analyse text with regard to reflection would greatly benefit research and practice.

Previous research has explored the potential of dictionary-based approaches that automatically map keywords to categories associated with reflection. Other automated methods use manually constructed rules to gauge insight from text. Machine learning has shown potential for classifying text with regard to reflection-related constructs. However, not much is known of whether machine learning can be used to reliably analyse text with regard to the categories of reflective writing models.

This thesis investigates the reliability of machine learning algorithms to detect reflective thinking in text. In particular, it studies whether text segments from student writings can be analysed automatically to detect the presence (or absence) of reflective writing model categories.

A synthesis of the models of reflective writing is performed to determine the categories frequently used to analyse reflective writing. For each of these categories, several machine learning algorithms are evaluated with regard to their ability to reliably detect reflective writing categories.

The evaluation finds that many of the categories can be predicted reliably. The automated method, however, does not achieve the same level of reliability as humans do.

ACKNOWLEDGMENTS

Writing this thesis has been a journey, in which I have met inspirational people all along my way. This is the place to thank them.

First and foremost my supervisors Dr. Fridolin Wild and Prof. Peter Scott. Both of them have been extremely talented in creating an environment in which I could flourish. From the beginning they were able to stimulate my own interest with research challenges that kept my learning curve high allowing me to build confidence in my own work. They were critical listeners helping me to pause were I would have rushed on in order to let me reflect and to find new ways of thinking about my research. They encouraged me with my work, supported me, guided me, and inspired me with their own work. I am grateful for their invaluable guidance.

This is also the place to recognise those wonderful people, without whom I probably would not have chosen an academic career. Looking back at my own pathway I want to thank Prof. Dr. Dr. Michael Henninger and Dr. Michael Balk, with whom I worked at the University of Regensburg at the department of Educational Science I - Analysis, Development and Evaluation of Learning Environments. They triggered my curiosity in educational research, while I was working for them as a student assistant. They gave me the opportunity to dive into the beauty of empirical research in the context of virtual learning and to fall in love with it.

I also want to thank Prof. Dr. Jörg M. Haake and Prof. Dr. Christoph Beierle from the FernUniversität Hagen. I see the course on CSCL/W Systems and the seminar on Knowledge Based Systems as the trigger for my own deep engagement with these topics and a door opener for many of my future activities.

My thanks go also to Dr. Christoph Hornung and Kawa Nazemi at Department A6 eLearning and Knowledge Management of the Institute for Computer Graphics Research of the Fraunhofer Society, Darmstadt. Working with them was truly inspiring and eye-opening for the importance of user models and adaptive visualisations.

This leads me to the Knowledge Media Institute of the Open University. It is a truly creative and stimulating environment for research. After coming from Germany to the UK, Dr. Viktoria Uren and Prof. Enrico Motta gave me my first opportunity to gain experience within the large scale EU research project X-Media.

KMI was not only an inspirational place for research but also a second home thanks to all of the amazing people I have met there. Thank you all for making it such a special place.

I am very thankful that I have met Prof. Simon Buckingham-Shum. Before my PhD started he engaged me with ELLI within the SocialLearn projects. ELLI is the Effective Lifelong-Learning Inventory. It brought me together with Prof. Ruth Deakin Crick from the University of Bristol and gave me the first taster into what great things can be achieved when educational scientists and computer scientists work closely together. He also fostered my interest in learning analytics and gave me several opportunities to work on interesting problems for SocialLearn and FutureLearn.

During my PHD I was working for the EU project STELLAR - the European Network of Excellence in TEL. This was a truly inspiring time and I am very grateful that I was able to work together with so many outstanding personalities of the technology-enhanced learning community.

Many thanks also to all my co-organisers of the Awareness and Reflection in Technology-Enhanced Learning workshop series and all the members of the programme committee and all the participants who made all four workshops a success.

A big thank you goes to Alba, who supported me all the time while I was working long hours and all those weekends. Keeping me in balance and reminding me that there is something else than working on this thesis.

Last but not least my thanks go to my family Renate, Joachim and Christian. I am so grateful for their enduring support and encouragement.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

1

INTRODUCTION

Online learning systems produce a staggering amount of text, such as essays, assignments, forum posts, comments, and feedback notes. Extreme cases are learning systems with massive participation, which pose challenges for current pedagogies (Ferguson and Sharples, 2014). Much of the thinking of students is expressed in their written contributions. The manual analysis of these writings is a laborious undertaking. Researchers would benefit from tools that automatically gauge insight from these textual contributions.

In recent years, the automated analysis of text has advanced significantly. Popular was the question-answering system Watson (Ferrucci et al., 2010) that was able to compete against human expert Jeopardy players by exploiting a large amount of structured and unstructured text. In the educational area, an example is the research on e-assessment (e.g., Kalz et al. (2014)) and research on automated essay assessments (Shermis and Burstein, 2003; Wild et al., 2005; Attali and Burstein, 2006; Alden Rivers et al., 2014; Jordan, 2014; Shermis, 2014) whose goal is to automatically grade student essays and provide feedback to students. Another strand of research, which is working on automated methods to derive meaning from text relevant for learning, is the broader field of applied natural language processing (for an overview, see McCarthy and Boonthum-Denecke (2012)) and discourse analysis (Dessus et al., 2009; Ferguson and Shum, 2011; Dascalu, 2014)).

The grand challenge reports for technology enhanced learning (TEL) indicated that 'e-assessment and automated feedback' is one of the highest ranked challenges (see

Sutherland et al. (2012, p. 21, 41 f.) and Fischer et al. (2014, p. 22ff.)). Sutherland et al. (2012, p. 21) stated that 'the grand challenge can be addressed by wide-scale development, evaluation and implementation of new formative assessment scenarios including the *development and evaluation of technologies that make for example intensive use of text- and data mining or natural language processing approaches* [emphasis added]'. Noss et al. (2012, p. 22) saw, in their vision for the UK education system, e-assessment as the opportunity for engaging students in activities by providing them and their teachers with *insights into their thinking*. These two TEL vision statements suggest that in order to tackle this grand challenge it is first necessary to develop the base technologies that can automatically provide these insights into thinking. We first need to develop reliable analysis technologies that then can be applied for assessment and automated feedback.

This research on the automated analysis of text for education is promising; however, insufficient research exists on automated methods analysing thinking skills expressed in writings (for a comprehensive overview of thinking skills see Moseley et al. (2004, 2005a,b)). Researchers have investigated the potential of automated analysis of critical thinking (McKlin 2004, p. 141; Corich 2011), a thinking type related to reflective thinking (Sparks-Langer and Colto 1991, p. 37,43; King and Kitchener 1994, chapter 1).

However, even less research exists on the automated analysis of reflective thinking in text. This is surprising because reflective thinking is core to educational practice. For example, the OECD report of Rychen and Salganik (2005, p. 8) saw reflection at the 'heart of key competencies' for 'a successful life and a well-functioning society'. The Assessment and Analytical Framework for PISA 2012 (OECD, 2013, p. 68) emphasised that reflection and evaluation are integral parts of reading literacy. The UK quality code for higher education recommends all UK higher education providers to ensure that all learning and teaching practices be informed by reflection, and recommends

a structured and supported process for learners to reflect upon their own personal development (QAA, 2012). The European commission funded two research projects whose core research topic was reflection from 2010 to 2014[1]. The meta-analysis of evidence-based practices in online learning of Means et al. (2010, p. 48), prepared for the U.S. Department of Education, highlighted the importance of providing online learning opportunities that promote reflection.

Reflective thinking is an important skill. In our current world, much of what people have to act on is unplanned and outside the usual routine. Practitioners can find themselves in ill-structured situations, for which learned standard recipes may not be immediately applicable. Such situations are full of uncertainty, they are unique, and values can conflict.

It is this scenario where the training of reflective thinking poses its value proposition. Reflective thinking can help find solutions to problems in situations that are highly undetermined. As Thorpe (2004) noted, '(...) educators seek to promote professional practice that is reflective rather than routine'. The key for this type of education is seen in helping students develop the ability to derive meaning from experience in order to better understand the reasons for their behaviour by considering past experiences, the present situation, and anticipated future results.

The absence of reflective thinking can lead to overconfidence and errors. Wald et al. (2012) concluded, based on findings in the area of health care, that 'failure to reflect on one's own thinking process, including critical examination of one's assumptions, beliefs, and conclusion, was recently described as a cognitive component of "physician overconfidence", a contributing cause of diagnostic error in medicine'.

The skills sought are the ability to inspect what occurred, critically analyse experiences in order to make informed decisions, and test their validity, which might refine or change one's practice. This type of thinking – reflective thinking – is seen as

1 See the projects MIRROR http://mirror-project.eu/ and ImREAL http://www.imreal-project.eu.

key for problem solving and professional practice (Schön, 1987) because it turns experience into learning (Boud et al., 1985).

Reflective thinkers challenge their thinking and acting routines in order to constantly improve practice. Given that reflection places emphasis on one's experiences, learning becomes a personal matter. Therefore, it is an act of active learning compared with the passive absorption of subject matter outside personal relevance.

It is this value proposition of reflection that gives it prominent position in educational research and educational practice.

An important educational practice to foster reflective thinking is reflective writing (for example, see Moon (2006) or Thorpe (2004)). Reflective writing is used as an educational tool in many disciplines that range from higher education, teachers' pre-service training, early childhood education, nursing, business, physical therapy, literature, psychology (Dyment and O'Connell, 2010, p. 234), and pharmacy (Wallman et al., 2008). Reflective writing aims at helping an individual to describe, re-think, and analyse their experience in order to reach a deeper understanding (Plack et al., 2005, p. 200). In addition, because the reflective thought process is made explicit in writing, it opens the possibility for learning through feedback and discussion on the reflective writing process (Bain et al., 2002, p. 193).

Reflective writing is an important educational practice and with it the analysis of reflective writing. Writing analysis is important for educators because it allows systematic assessment of the quality of the reflective writing or writer, and it is important for researchers because it allows to investigate the influence of specific instructions on the improvement of the reflective writing skills.

The prevalent technique to assess reflective writing is the application of principles of content analysis. Although it is important, several researchers indicated that research on assessing reflection in writing is not fully developed. In their review of content analysis procedures used to assess reflective writings, Poldner et al. (2012, p.

32) concluded that 'researchers have only recently begun to develop quality assessment criteria'. Wong et al. (1995, p. 49) noted a lack of empirical research on assessing reflection. A similar sentiment was expressed by Plack et al. (2005, p. 199), '(...) yet little is written about how to assess reflection in journals'.

There is a need to advance the methods used to analyse reflective writing. The aim of this thesis is to explore the potential for automatizing this manual content coding process using machine learning techniques. The automated detection of reflection promises advancement in research and practice, especially because it would add another method to the analytical repertoire of researchers, and it has the potential of supporting educators with their assessment of reflective writing.

Considering the importance of reflective writing to foster one's reflective thinking, the automated analysis of such writing would be a promising technology to support researchers, teachers, and learners. Advancement of automated text analysis suggests the feasibility of automating the analysis of reflection in text. However, there is insufficient research on analysing reflection automatically. This thesis sets out to contribute in closing this gap by researching the potential of machine learning algorithms to detect reflection in text.

1.1 RESEARCH QUESTIONS

The goal of this research is to evaluate the reliability of automated methods to analyse reflection in student writings. In this thesis, the investigated automated method is machine learning. Not much is known of whether machine learning can be used to detect reflection. Detection in this context means that the automated method can determine whether a given text segment is reflective. Parallels can be drawn between the manual content analysis of reflective writing and the intended automated process.

As with the manual coding process, which assigns labels to analysis units, the automated reflection detector provides these labels automatically.

In detail, **this thesis investigates whether text segments can be analysed automatically using machine learning algorithms to detect the presence (or absence) of reflection**.

Content analysis is one of the main methods for assessing written reflection. Text analysis is guided by a model that describes the coding categories. Several models of reflective writing have been proposed. Although these models vary in their description of model constituents, two characteristics reoccur frequently. The first is that a writing can be categorised with regard to its level of reflection. Models that describe reflection levels have in common that they order writing from descriptive/non-reflective to reflective. The second characteristic describes categories associated with reflection. Examples of such categories are the description of experience, awareness of problems, evidence of critical analysis, etc.

This thesis investigates the automated detectability of both aspects of reflection. These aspects are encapsulated two research questions.

The first research question is: **Q1: Can machine learning algorithms be used to distinguish between descriptive and reflective text segments?**

This question addresses the level aspect of reflective writing. The common denominator of all level models of reflective writing is the distinction of text as descriptive/non-reflective or reflective. The aim here is to investigate whether machine learning can be used to discern text segments regarded as reflective from descriptive text segments.

In order to structure the argument for (or against) the first research question, three lines of investigation are proposed. Once answered, they provide evidence in favour of the main research question (or against it). The three lines of investigation are:

- **I1: Can tree-based machine learning algorithms detect the difference between descriptive and reflective texts segments?**
- **I2: Can rule-based machine learning algorithms detect the difference between descriptive and reflective text segments?**
- **I3: Can high performance machine learning algorithms detect the difference between descriptive and reflective text segments?**

Within each line of investigation, a set of machine learning algorithms is evaluated on the problem of detecting reflection. Because there is no prior information available on which machine learning algorithm would perform well on this specific problem, several machine learning algorithms were chosen to evaluate their potential for the automated detection of reflection. Each line of investigation contains a set of specific machine learning algorithms selected with regard to their potential for detecting reflection (see Section 4.5 'Machine learning algorithms').

Tree-based machine learning algorithms automatically generate decision trees from sample data to classify text.

Rule-based machine learning algorithms construct rules from data. Rules follow the common pattern of defining premises that, when satisfied, allow deriving conclusions automatically.

Both tree-based and rule-based algorithms have the property of producing models that can be inspected in the form of decision trees or rule sets.

The third line of investigation summarises the machine learning algorithms that tend to perform well on many tasks and often outperform tree-based and rule-based machine learning algorithms. The investigated machine learning algorithms are: Support Vector Machines (SVM), Neural Networks, Naïve Bayes, and Random Forests. Here, I refer to these different classes of machine learning algorithms as **high performance machine learning algorithms**.

All selected machine learning algorithms have shown potential for text classification tasks. However, it is not known which performs well on the task of reflection detection. The three lines of investigation provide a comparative overview of which candidate machine learning algorithms has the highest performance for predicting reflection.

The second research question addresses the second aspect of the reflective writing models. These are the categories associated with reflection. The reflective writing models vary with regard to these categories. Notwithstanding their differences, some of the categories occur frequently. They are: description of an experience, awareness of feelings, awareness of one's personal perspective, having a critical stance, considering other perspectives, and the description of outcomes. In this thesis, they are referred to as common categories of reflective writing (see Section 2.3 'Model for reflection detection').

Therefore, the second research question is: **Q2: Can machine learning algorithms be used to detect common categories of reflective writing?**

This question explores the reliability of the automated detection of concepts that are common in reflective writing. As with the first research question, a set of machine learning algorithms is used to assess the detectability of the common categories of reflection.

Both research questions contribute to the wider goal of this thesis. The first question addresses the detectability of reflective and descriptive text segments. The second question explores the detectability of categories that are common in reflective writing. If it can be shown that it is possible to automatically detect both aspects, a strong case can be made for the applicability of machine learning methods to detect reflection in text.

1.2 THESIS OVERVIEW

This thesis consists of seven main building blocks.

Chapter 1 'INTRODUCTION' provides the context of this research. It highlights the importance of reflection for our society and sets out the goal of this thesis to research machine learning methods in order to automate the detection of reflection in text. The research questions set the specific aim of this investigation.

Chapter 2 'THE CONCEPT OF REFLECTION: THEORY AND MODEL' introduces the concept of reflection starting with its use in everyday language, highlights its several meanings, and provides its central definitions. Then, the chapter provides a comprehensive overview of the models of written reflection. These are the models used to analyse reflection in writing with principles of the content analysis. A synthesis derives those categories that are common for all models. The result of this synthesis is a model of common categories of reflective writing. In the evaluation, each category forms a test case with regard to its detectability with machine learning algorithms.

Chapter 3 'RELATED METHODS AND BENCHMARKS' provides an overview of the main method for analysing reflective writing and automated methods that have been applied to detect related thinking skills. The main method to assess written reflection is content analysis. After outlining the core principle of the content analysis, two of the content analysis-specific practices are described. The first strategy concerns the relationship between the annotated analysis units and the assignment of a category to the entire text, and the second strategy describes the practice of mapping descriptive categories to levels of reflection. Subsequently, a comprehensive overview of rater performance in analysing written reflection follows. The second part of this chapter informs on the automated methods used to gauge insight from text on the reflection

aspects. As with the part on the manual analysis of reflective writing, the part on automated methods summarises performance measures that provide insight on what can be expected from automated methods with regard to the reliable detection of model categories.

Chapter 4 'METHODOLOGY AND RESEARCH DESIGN' outlines the methodological considerations that shaped the research design. This chapter provides the rationale for the type of study, and gradually introduces evaluation criteria and relevant measurements, as well as discusses analysis units and sampling techniques. Subsequently, the chapter outlines the decisions that led to the selection of those machine learning algorithms used in the evaluation. These considerations shaped the research design. The research design section outlines the data generation process and the design of the evaluation of the machine learning algorithms.

Chapter 5 'DATASET GENERATION' outlines the specifics of the data generation process. It starts with a large collection of text and ends in several datasets suitable for evaluating the machine learning algorithms on the problem of reflection detection. Each section is a step of the transformation process outlined in the data generation research design. The chapter describes how the text collection was identified, which texts were selected, how the texts were divided into smaller text segments, and how each of the text segments was annotated with regard to indicators of the reflection categories.

Chapter 6 'EVALUATION' implements the steps outlined in the research design and provides a detailed evaluation of the performance of machine learning algorithms on the problem of reflection detection. The chapter is divided into two major parts aligned to the two main research questions. The first part provides an evaluation of the machine learning algorithms to detect reflection. It reports the performance of the machine learning algorithms following the three lines of investigation. The second part reports the findings on the performance of machine learning algorithms to detect each of the

indicators of the common categories of reflection. The performance measures derived in the evaluation part of this thesis are the basis for the argument in favour (or against) the reliable detection of reflection with machine learning.

Finally, Chapter 7 'CONCLUSION AND FUTURE RESEARCH' draws conclusions regarding the research questions and describes future research.

2

THE CONCEPT OF REFLECTION: THEORY AND MODEL

For the automated analysis of reflection, it is important to clearly understand what constitutes reflection. The following sections provide an overview of the meaning of reflection, starting with its use in everyday language and ending in its specific use in educational discourse.

Reflection is widely used in everyday language. We talk about reflection using phrases like 'when I reflect back on my last year', 'I never realised this about myself', or 'if only I had known then what I know now'. In everyday language, reflection has many meanings. For example, it is sometimes used as a synonym for general thoughts or a collection of ideas about something, sometimes it is used as a synonym for critical thinking or thinking about a problem. And it is used in the sense of a long and deep thought about an experience.

To start delving into the meaning of 'reflection', a look at its origin can be of help. The word 'reflection' has its roots in the late Latin word *reflexio* – the act of bending back[1], which originates in the in the Latin verb *reflectere*. The word reflection has different meanings, ranging from 'a calm, lengthy, intended consideration', 'the phenomenon of a propagating wave (...) being thrown back from a surface', to 'the image of something as reflected by a mirror'[2]. They all, however, share the meaning of its Latin root: the act of bending back. In the case of the 'calm, lengthy, intended consideration', the act of bending back can be associated with directing attention towards a past experience (bending one's thought back towards something). A wave that is 'thrown back from a

1 http://www.merriam-webster.com/dictionary/reflection and http://www.oxforddictionaries.com/definition/english/reflection

2 http://wordnetweb.princeton.edu/perl/webwn?s=reflection

surface' is bent at an angle. The same applies to mirror reflection: Light is 'bent' at an angle when being reflected by the mirror.

The act of bending one's attention towards an important experience, the mirror allowing us to see ourselves, and all the different angles from which a situation can be interpreted are all important associations with the concept of reflection as it is used in educational discourse (Ratkic, 2012).

2.1 DEFINITIONS OF REFLECTION

The following definitions will help to determine what is reflection and what qualities it has. The definitions are chosen from high impact research articles. The impact was determined by a co-citation analysis (see Appendix A for the details of this analysis). These are the definitions of Dewey (1933), Schön (1983, 1987), Boud et al. (1985, chapter 1), and Mezirow (1991).

Definitions aim to describe the essence of a concept, and thus may appear as being overly simplified when presented without their original context. It is therefore important to bear in mind that reflection is a complex construct and any description of it will fall short of describing its richness (Jay and Johnson, 2002).

Frequently, Dewey is mentioned as the person who brought reflective thinking into the educational discourse. Dewey (1933) defined reflection as an 'active, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it, and the further conclusion to which it tends.' Reflection in Dewey's sense means testing and challenging validity (Mezirow, 1991).

The work of Schön (1983, 1987) is often mentioned, because of the introduction of the concept of reflective practice. He is often referred to when talking about the distinction between reflection-in-action and reflection-on-action. Reflection-in-action is exemplified by practitioners who, while facing a problem, often have to stop their

routine, think about the problem, and solve it *in situ*. Reflection-on-action happens after the event. Practitioners return or revisit their experience in order to shed light on the problem from alternative perspectives. Several other distinctions have been proposed. Less frequently cited distinctions can be found in Killion and Todnem (1991) about 'reflection-for-action', Greenwood (1993, p. 1186) about 'reflection-before-action', and Chrzaszcz et al. (2008) about 'reflection-on-reflection'. Eraut (1995) proposed a re-conceptualisation of 'reflection-in-action' and 'reflection-on-action'.

Boud et al. (1985, chapter 1) saw reflection as: 'An important human activity in which people recapture their experience, think about it, mull it over and evaluate it.' Reflection involves both cognitive and affective engagement, which leads to new insights and understanding of experience (Boud et al., 1985). Boud et al. saw reflection not as a cognitive act alone. In this theory, emotions play a significant role (Boud et al., 1985, p. 11).

Mezirow (1991, p. 101) emphasised the function of reflection in learning. Reflection '(...) makes enlightened action and reinterpretation possible, and especially for the crucial role that reflection plays in validating what has been learned.' According to Mezirow (1990a, p. xvi), reflection has three functions: to guide action, to give coherence to the unfamiliar, and to reassess the justification for what is already known. Mezirow (1991, chapter 4) distinguished three qualities of reflective thinking: content, process, and premise. Content reflection addresses the question as to '(...) what we perceive, think, feel, or act upon' (Mezirow, 1991, chapter 4). Process reflection investigates 'how we perform these functions of perceiving, thinking, feeling, or acting and an assessment of our efficacy in performing them' (Mezirow, 1991, chapter 4). Premise reflection involves questioning '(...) why we perceive, think, feel, or act as we do and of the reasons for and consequences of our possible habits of

hasty judgements, conceptual inadequacy, or error in the process of judging (...)' (Mezirow, 1991, chapter 4).

The definitions of reflection should provide a guiding structure about what is understood by reflective thinking. This section should help us to understand that there is no such thing as the definition of reflection, but that there is an on-going discourse about what constitutes reflection. Many of the reflective writing models are based on the definitions of reflection. The next section provides an overview of reflection models. The common constituents of these models will inform the model for reflection detection.

2.2 MODELS TO ANALYSE WRITTEN REFLECTION

Several models of reflection in the context of the analysis of writings have been proposed. Models of reflection characterise the components of reflection and their relations with each other. Conceptualisations of reflection are the first step towards empirical research of reflection. Models provide an abstraction of the phenomenon of reflection. They represent a simplified version of reflection. Yet, they are detailed enough to study and test assumptions about reflection with empirical methods. Researchers sometimes refer to these models as frameworks (Fund et al., 2002; Ward and McCotter, 2004), rubrics (Ward and McCotter, 2004; Wald et al., 2012) or express them in a concrete coding schema (Kember et al., 1999; Poldner et al., 2014). Here, we will refer to them as models, as they characterise and operationalise qualities of reflection found in writings (examples of such models can be found in Table 1).

Many models have been proposed. There exists a wide variety of theory-informed models and models that have been derived by qualitative research. Examples thereof can be found in the work of Ross (1989), Sparks-Langer and Colto (1991), Gore and Zeichner (1991), Tsangaridou and O'Sullivan (1994), Hatton and Smith (1995),

Richardson and Maltby (1995), Pultorak (1996), Hutchinson and Allen (1997), Scanlan and Chernomas (1997), Taylor (1997), Valli (1997), Bain et al. (1999), Kim (1999), Duke and Appleton (2000), Rogers (2001), Bain et al. (2002), Jay and Johnson (2002), Spalding et al. (2002), MacLellan (2004), Tillema (2004), Thorpe (2004), Ward and McCotter (2004), Lee (2005), Korthagen and Vasalos (2005), Kansanaho et al. (2005), Kreber (2005), Wessel and Larin (2006), Mann et al. (2007), Chretien et al. (2008), Kreber and Castleden (2008), Minott (2008), Wilson (2008), Gulwadi (2009), Friedman and Schoen (2009), Le Cornu (2009), Badger (2010), Granberg (2010), Lambe (2011), Cohen-Sayag and Fischl (2012), Crawford et al. (2012), Etscheidt et al. (2012), Leijen et al. (2012), Corlett (2013), Medwell and Wray (2014), McDonald et al. (2014), Nguyen et al. (2014), Chaumba (2015), Hill et al. (2015), and McKay and Dunn (2015).

The following literature review focusses on models that have been empirically evaluated and are based in the area of the analysis of reflective writings (see Table 1). Specifically, papers were selected only if they contained a statement about the reliability of the coding process.

The rationale for this inclusion criterion is that the research on the empirically evaluated analysis of reflective writing shows close parallels to this investigation. The empirically evaluated analysis of reflective writing is driven by the method of content analysis. Each text or text segment is assessed by a number of raters and labelled with a category of the chosen reflection model. The quality of the coding process was evaluated by estimating the reliability of the coding process. Similarly is the aim of this thesis because its goal is to automatically label text segments with a category of the model for reflection detection and estimate the reliability of the coding process. Reliability measures inform on the quality of the coding process, and therefore, are central for answering the research questions (see Section 1.1 'Research questions'). The evaluation criteria and measurements, including reliability, are discussed later in Section 4.2 'Evaluation criteria and metrics'. The reported reliabilities of the manual

analysis of reflective writing can provide guidance on the reliability that can be expected from the content analysis of reflective writing. This should allow discussing the reliability of automated methods in the context of the reliability of the manual coding process.

The studies outlined above do not provide this information, and therefore, do not allow inferring information on the degree to which codes can be reliably assigned to text.

The models that satisfied these inclusion criteria are listed in Table 1. This list was the result of a thorough and repeated search of literature databases over several years, including a systematic analysis of the literature cited by the retrieved papers. Some key papers, which contained similar lists, but tailored to the aims of the authors of these publications, can be found in Dyment and O'Connell (2011, p. 85 f.), Birney (2012, p. 90 f.), Poldner et al. (2012, p. 33 ff.), and Poldner et al. (2014, p. 351 ff.).

These empirically-grounded models used for the analysis of reflective writing are often closely informed by the theory of reflection and thus cover a wide range of well-established theoretical models (see the column 'based on' in Table 1 and the referenced literature). Most of the research on the analysis of reflection in writings can be traced back to the theoretical ideas of Boud et al. (1985, chapter 1) (e.g. Wong et al. (1995); Wald et al. (2012); Plack et al. (2005)) and Mezirow (1990a) (e.g. Wong et al. (1995); Kember et al. (1999); Plack et al. (2005); Wald et al. (2012)), the ideas on reflection of Manen (1977) (e.g. Sparks-Langer et al. (1990) and Poldner et al. (2014)[3]), and the work of Schön (Schön, 1983, 1987) (e.g. Plack et al. (2005); Wald et al. (2012)). While these models stem directly from the theoretical work on reflection, there are also models using conceptualisations of learning outcomes based on the learning taxonomy of Bloom (1954) (e.g. Plack et al. (2007); O'Connell and Dyment (2004)).

3 The model in Poldner et al. (2014) is based on Leijen et al. (2012), which in turn is based on Van Manen (1977).

This information about which authors and with it which theoretical ideas were influential for the models is helpful to derive a clustering of the school of thought underlying these models. Schools of thought (also called paradigms or perspectives), such as constructivism, cognitivism, or behaviourism, summarise a set of beliefs assumed to be true about the world (Schuh and Barab, 2008, p. 71). Making these assumptions explicit can help to better understand these models. Here, we restrict the discussion to these three mentioned perspectives (see for an extensive overview of schools of thought Schuh and Barab (2008); Ertmer and Newby (2013); van Merriënboer and de Bruin (2014)).

On the one hand, one can argue that the models from authors such as Boud et al., Mezirow, Manen, and Schön, have a constructivist influence as these authors oriented their work on the ideas of Dewey about reflection[4]. Dewey's positions are often associated with one of the constructivist perspectives (see Garrison (1995, p. 717); Bond (2003, p. 10); Anderson and Dron (2010, p. 84)). Fenwick (2001, p. 17ff.) also attested these authors a constructivist perspective.

On the other hand, the work of Bloom is associated with the cognitivist paradigm (Bates, 2015, p. 49). Therefore, models that are largely influenced by the learning outcome taxonomy of Bloom (1954) can be said to be influenced by cognitivist assumptions.

The third big perspective, behaviourism, is generally not associated with the perspectives of above outlined authors on which the models are based on. This school of thought focusses only on observable behaviour and not on inner thought processes (Schuh and Barab (2008, p. 73); Ertmer and Newby (2013, p. 48 ff.); van Merriënboer and de Bruin (2014, p. 24f.)) as reflective thinking expressed in writings.

4 Boud et al. (1985, p. 21) wrote that they '(...) acknowledge a great debt to Dewey (...)'. Mezirow (1990a, p. 10off.) built on Dewey's view of reflection, similar Van Manen (1995). Schön wrote his PhD thesis about Dewey (Schön, 1987, p. xi).

The paradigm that explicitly focusses on internal cognitive processes is cognitivism. Compared to the behaviouristic perspective learning is seen as '(...) changes between states of knowledge rather than with changes in the probablity of responses' (Ertmer and Newby, 2013, p. 51). Both, the behaviouristic and cognitivistic view originate in objectivism; i.e. the assumption that the world exists/is real (Schuh and Barab (2008, p. 73)). Instruction under behaviouristic and cognitivist perspective means '(...) to map the structure of the world onto the learner' (Ertmer and Newby, 2013, p. 54).

The (cognitive) constructivist perspective acknowledges the existence of reality, but reality is seen as an interpretation and that this interpretation is shaped by experience of an individual (Schuh and Barab, 2008, p. 74). Meaning is individually constructed. Another variant of constructivism, social constructivism, sees this meaning making process as a social rather than an individual experience (Schuh and Barab, 2008, p. 74).

In summary, the models that have been considered as relevant for this thesis (see Table 1), can be clustered according to the two schools of thought namely cognitivism and constructivism. This result aligns with the observation of Birney (2012, p. 16), which noted that some researchers determined reflective learning as a cognitivist approach, while others were seeing it as a constructivist approach.

Table 1 lists the model descriptions, along with author information and the research on which the model was based.

Author	Based on	Model
Sparks-Langer et al. (1990)	Gagne (1968) hierarchy of thinking and Van Manen (1977) idea of critical reflection	Levels: 1. No descriptive language, 2. Simple, layperson's terms, 3. Events labelled with appropriate terms, 4. Explanation with tradition or personal preferences given as rationale, 5. Explanation with principles or theory given as the rationale, 6. Explanation with principles/theory and consideration of contextual factors, 7. Explanation with consideration of ethical, moral, political issues
Wong et al. (1995)	Elements: Boud et al. (1985, chapter 1). Levels: Mezirow (1990a)	Elements: Attending to feelings, association, integration, validation, appropriation, outcome of reflection. Levels: non-reflector, reflector, critical reflector

Continued on next page

Table 1 Continued from previous page

Author	Based on	Model
Sumsion and Fleet (1996)	Surbeck et al. (1991) and Boud et al. (1985)	Final instrument was not shown. They refer to Surbeck et al. (1991) and Boud et al. (1985). Model of Surbeck et al. (1991) had the three categories: reaction, elaboration, and contemplation. Each category has several subcategories. Boud et al. (1985, p. 27–35) used the following stages: 1. Returning to experience; 2. Attending to feelings (utilizing positive feelings, removing obstructive feelings); 3. Re-evaluation of experience (association, integration, validation, appropriation). Outcome. Sumsion and Fleet (1996) used the terms: Not reflective, moderately reflective, highly reflective
McCollum (1997)	Tsangaridou and O'Sullivan (1994)	Levels: Description, description and justification, description and critique, description, justification and critique. Focus: Technical, situational, sensitizing
Kember et al. (1999)	Mezirow (1991)	Non-reflective: Lower level: 1. Habitual action, higher level: 2. Introspection and 3. Thoughtful action. Reflective: Lower level: 4. Content reflection, 5. Process reflection, 6. Content and process reflection, higher level: 7. Premise reflection
Hawkes and Romiszowski (2001); Hawkes (2001, 2006)	Sparks-Langer et al. (1990)	Levels: 1. No description of event; 2. Events and experiences, described in simple, layperson's terms; 3. Description of events and experiences employing pedagogical terms; 4. Explanation of events or experiences is accompanied by rationale or tradition or personal preference; 5. Explanation of an event or experience using cause/effect principle; 6. Explanation provided that identifies cause and effect factors while also considering contextual factors; 7. Explanation of events, experience, or opinion that cites guiding principles and current context, while referencing moral and ethical issues
Fund et al. (2002)	Form of writing is based on Hatton and Smith (1995)	WRITT evaluative tool: 1st dimension (object or content of writing): Subject-matter content; didactic content; personal content. 2nd dimension (form of writing): Lower-level reflection: Description, personal opinion, Higher-level reflection: Linking, critical bridging
Hamann (2002)	LaBoskey 1994	Commonsense thinker (unreflective): a) Self-orientation, b) short-term view, c) reliance on personal experience in learning to teach, d) metaphor of teacher as transmitter, 3) lack of awareness of need to learn, f) overly certain conclusions, g) broad generalizations, h) existing structures taken as given, i) lack of commitment, j) views classroom in isolation. Alert novice thinker (reflective): k) Student orientation, l) long-term view, m) differentiation of roles of the teacher and learner, n) metaphor of teacher as facilitator, o) openness to learning, p) acknowledgement of need for conclusions to be tentative, q1) means–end thinking, q2) reasoning grounded in knowledge of self, r) imaginative thinking, s) awareness of teaching as a moral activity, awareness of the classroom and teaching as part of a social and political context
Pee et al. (2002)	Hatton and Smith (1995)	Types of writing: 1. Descriptive; 2. Descriptive reflection; 3. Dialogic reflection; 4. Critical reflection
Williams (2000) cited in Williams et al. (2002)	Boud et al. (1985, chapter 1)	Levels: 1. Describes learning. 2. Analyses learning. 3. Verifies learning. 4. Gains new understanding. 5. Indicates future behaviour. Non-reflective journals: Is descriptive in nature, reporting on what is happening rather than analysing the learning event, issue, or situation
Boenink et al. (2004)	Own invention	Scores: 1–2 Oversimplified, intolerant opinion, only emotional reaction. 3–4 Limited/restricted, narrow-viewed, one-sided reaction, mostly just 1 perspective, no weighing up or balancing, no attention paid to context. 5 More than 1 perspective, but neither balancing nor attention paid to context. 6–7 More perspectives, general as well as personal, some balancing between perspectives. 8–9 Differentiated balancing, room for dilemmas and or doubt, explicit attention paid to the patient. 10 A subtle/balanced approach, considering all relevant perspectives, weighing up of different interests, a keen eye for dilemmas and uncertainties, paying attention to the patient's viewpoint and an evaluation of one's own position and latitude
O'Connell and Dyment (2004)	Bloom (1956)	Three categories: a) Type, b) Bloom's Taxonomy of Cognitive Thinking, c) creative entries. Type contained amongst others: Factual information, personal reflection, transfer. Bloom's Taxonomy: Knowledge, comprehension, application, analysis, synthesis, evaluation
Plack et al. (2005)	Boud et al. (1985, chapter 1), Mezirow (1990a), Schön (1987)	Components of reflection: Reflection in action, reflection on action, reflection for action, content reflection, process reflection, premise reflection, returns to experience, attends to feelings, and evaluation of experience. Levels: no evidence of reflection, evidence of reflection, evidence of critical reflection
Ballard (2006)	Mezirow (1991)	Levels: Technical rationality, practical action, critical reflection

Continued on next page

Table 1 Continued from previous page

Author	Based on	Model
Mansvelder-Longayroux (2006); Mansvelder-Longayroux et al. (2007)	Vermunt and Verloop (1999)	Activities: Recollection, evaluation, analysis, critical processing, diagnosis, reflection
Abou Baker El-Dib (2007)	Various inspirations and Kember et al. (1999)	Stages: Statement of problem, plan of action, acting, reviewing. Levels: Low, low medium, high medium, high
Chirema (2007)	Elements: Boud et al. (1985, chapter 1). Levels: Mezirow (1990a), 1991	Elements: Attending to feelings, association, integration, validation, appropriation, outcome of reflection. Levels: non-reflector, reflector, critical reflector
Plack et al. (2007)	Bloom (1956)	Levels: 1. Knowledge and comprehension (data gathering); 2. Analysis (data analysis); 3. Synthesis and Evaluation (conclusion drawing)
Kember et al. (2008)	Mezirow (1991)	Categories: Non-reflection, understanding, reflection, critical reflection
Wallman et al. (2008)	Kember et al. (1999) and Mezirow (1991)	Non-reflective: 1. Habitual action; 2. Thoughful action; 3. Introspection. Reflective: 4. Content reflection, 5. Process reflection; 6. Premise reflection
Chamoso and Cáceres (2009)	Various inspirations	Categories: Control, generality, description, argumentation, contribution
Findlay et al. (2010)	Based on Boud et al. (1985, chapter 1) and Wong et al. (1995) (see Findlay et al. (2009))	Deep analytic levels of the Newcastle Reflective Analysis Tool (NRAT): No evidence of reflection; Level 1: Attending to feelings; Level 2: Association; Level 3: Integration; Level 4: Validation; Level 5: Appropriation; Level 6: Outcome of reflection. Broad classification NRAT: No evidence of reflection -> non reflector; levels 1–3 -> reflector; levels 4–6: critical reflector
Lai and Calandra (2010)	Ward and McCotter (2004); Levels mirror Van Manen (1977) stages	Rubric: Levels: Routine, technical, dialogic inquiry, transformative. Dimension 2: Focus, inquiry, change
Bell et al. (2011)	Kember et al. (1999)	Categories: Non-reflective: Introspection, thoughtful action. Reflective: Content reflection, process reflection - internal, process reflection - others, process reflection - others/internal, content reflection/process reflection - internal, content reflection/process reflection - others, premise reflection
Clarkeburn and Kettula (2011)	Kember (2008)	Levels: Habitual action (non-reflection), understanding, transitional, reflection, critical reflection
Findlay et al. (2011)	Based on Boud et al. (1985, chapter 1) and Wong et al. (1995) (see Findlay et al. (2009))	See Findlay et al. (2010)
Fischer et al. (2011)	Levels: Mezirow (1991)	Non-reflective, reflection on experience low level, reflection on experience high level, reflection on awareness
Birney (2012)	Results of a Delphi study	Indicators: 1. Clear description of context, 2. Issues are identified, 3. Analysis is evident, 4. Creative synthesis is evident, 5. Implications of actions are considered, 6. Multiple perspectives are considered, 7. Links are made to broader social structures, 8. Learning is evident, 9. Insightful understanding is evident, 10. Changes in beliefs or understanding are evident, 11. Revisions to future practice are discussed, 12. Self-awareness is evident. Levels: Descriptive (indicator 1); low-level reflection (indicator 2–4); medium-level reflection (indicator 5–7); high-level reflection (indicator 8–12)
Ip et al. (2012)	Wong et al. (1995)	Levels: Non-reflector, reflector, critical reflector
Wald et al. (2012)	Schön (1983), Boud et al. (1985, chapter 1), Moon (1999), and Mezirow (1991)	Writing spectrum (descriptive to critique of beliefs), presence of the author, description of conflict or disorienting dilemma, considering emotions, analysis and meaning-making, attendance to assignment (optionally). Levels: habitual action, thoughtful action, reflection, and critical reflection. Plus transformative and confirmatory learning as outcome at the level of critical reflection
Mena-Marcos et al. (2013)	Levels based on Hatton and Smith (1995). Type of reflection based on Tillema (2004)	Levels: Simple or habitual reflection, descriptive reflection, dialogical reflection, critical reflection. Types: Appraisal (either positive or negative), rules (deliberate if–then statements or conceptual guidelines), artefacts (as a potential solution to a problem)

Continued on next page

Table 1 Continued from previous page

Author	Based on	Model
Poom-Valickis and Mathews (2013)	Hatton and Smith (1995)	Types: 1. Unreflective, descriptive writing. 2. Descriptive reflection. 3. Dialogic reflection. 4. Critical reflection
Poldner et al. (2014)	Leijen et al. (2012)	Categories: Description, evaluation, justification, dialogue, transfer. Each category had several subcategories
Prilla and Renner (2014)	Phases: Own coding schema. Levels: Based on Fleck and Fitzpatrick (2010) and de Groot et al. (2014)	Phases/codes: 1. Description of an experience. 2. Mentioning and describing emotions. 3. Interpreting or explaining behavior in the experience. 4. Linking an experience explicitly to other experiences. 5. Linking an experience to knowledge. 6a. Responding to the explanation of an experience by providing alternative perspectives. 6b. Responding to the explanation of an experience by challenging or supporting assumptions. 7a. Contributing to work on a solution by providing reason for the issue. 7b. Contributing to work on a solution by providing solution proposals. 8a. Showing insights or learning from reflection by describing better individual understanding. 8b. Showing insights or learning from reflection by generalising from reflection. 9. Describing or implementing change. Stages: 1. Provision and description of experience (codes 1, 2). 2. Reflection on experience (codes 3–7). Learning or change (codes 8, 9)

Table 1: Models of reflection

From the model column in Table 1 it can be seen that authors characterise the constituents of their model with words like level, element, dimension, type, score, categories, components, activities, stages, indicators, spectrum, phase, or code. Although those are differences, these models, by and large, exhibit two main qualities, which can be described as the quality of depth and breadth (Moon, 2004, p. 95 ff.).

Many models involve the quality of depth. These are models that conceptualise the quality of reflection as a hierarchy with levels that range from non-reflective up to highly reflective. An example is the model of Ip et al. (2012), which ranges from non-reflector to critical reflector. Another example is the model of Kember et al. (1999), which spans from the non-reflective level of habitual action to the highest form of reflection, called premise reflection. And a third example is the model of Lai and Calandra (2010), which ranges from routine (non-reflective) up to transformative (highly reflective). Other examples are the model of Sparks-Langer et al. (1990) , McCollum (1997), or Ballard (2006). Level models often refer to the work of Manen (1977, p 226 f.), who spoke of levels of reflectivity, and Mezirow (1991, chapter 4), who contrasted non-reflective action with reflective action. Mezirow distinguished three types of non-reflective action: habitual action, thoughtful action, and introspection;

and three types of reflective action: reflection about content, about process, and about the premises (see Section 2.1 'Definitions of reflection'). Mezirow (1991, chapter 4) saw premise reflection as the highest form of reflection, as it can lead to a change of perspective.

Most models contain the breadth (descriptive) dimension of reflection. This dimension of reflection describes types of reflection emphasising the individual usefulness of each category as a distinct characteristic of reflective writing, without implying a hierarchical structure as does the level dimension. Wong et al. (1995), for example, describes reflection using categories such as 'attending to feelings', 'validation,' and 'outcome of reflection'. Poldner et al. (2014) used the categories 'description', 'evaluation', justification', 'dialogue', and 'transfer'. Examples of this type can be found in Wong et al. (1995), Wald et al. (2012), Birney (2012), Mansvelder-Longayroux (2006) and Mansvelder-Longayroux et al. (2007).

In the context of the content analysis of reflective writings, these descriptive categories are the coding categories usually applied to parts of the text, while the depth dimension is usually used to categorise the whole text.

These two qualities are not the only two qualities for describing models of reflection. They occur, however, very frequently in work analysing written reflection, as shown in Table 1. To give an example of another quality, I resort to more theoretical work. There, one can find models of reflection that describe the iterative, cyclic nature of reflection (see also Mann et al. (2007, p. 598)). These are models which emphasise the process of reflection by describing the prototypical flow of reflection. Examples are the models of Scanlan and Chernomas (1997), Schön (1987, p. 27ff.), Moon (2004, p. 214 f.), and Korthagen and Vasalos (2005, p. 57 f.). This iterative or process nature of reflection is, however, not the focus of the examined work analysing written reflection as summarised in Table 1, nor is it the focus of this thesis, and thus will not be further elaborated on.

Many of the models listed in Table 1, which contain both the breadth and depth qualities of reflection, describe a mapping mechanism between their categories and levels of reflection. Often, this mechanism maps one or several categories of reflection to a level of reflection. Once the categories of reflection are determined, the chosen mapping approach assigns a level to the text. For example, Birney (2012, p. 214) assigned texts to the high-level of reflection if the text showed evidence of the categories 'learning', 'insightful understanding', 'changes in beliefs', 'revision of future practice', and 'self-awareness'. These mapping mechanisms are outlined in detail in Section 3.1.3 'Relationship between the descriptive and level reflection quality'. It is notable that it is necessary to first determine the descriptive (breadth) categories of reflection in order to then map them to levels. The determination of the level is solely based on the descriptive categories. This insight led this research to focus especially at the descriptive categories of reflection. A close look at this breadth dimension of reflection follows in Section 2.3 'Model for reflection detection'.

Next, three models have been chosen to illustrate the character of models of reflective writing. These three models have been frequently mentioned in other work (see column 'based on'). They will serve to highlight important features of reflective models. They are the models of Wong et al. (1995), Hatton and Smith (1995), and Kember et al. (1999) (see Table 1). The models cited in Table 1 will be discussed again in Section 2.3 'Model for reflection detection'.

The model of Wong et al. (1995, p. 57) consisted of six elements of reflection, which were derived from the work of Boud et al. (1985). They are the seven elements 'attending to feelings', 'association', 'integration', 'validation', 'appropriation', and 'outcome of reflection'. For each element, they described a set of coding criteria. For example, a text segment was labelled as 'integration' if it showed evidence that prior knowledge was linked to new knowledge, but also if the text described that the writer had a new insight. Text was labelled as 'attending to feelings' if there was evidence

that the writer described either the intention to overcome negative feelings or to make use of positive feelings. For each element Wong et al. (1995, p. 52 ff.) provided text snippets with an explanation for the coding decision. These categories were then mapped to three levels: Non-reflector, reflector, and critical reflector. A non-reflector did not show any evidence of any of the elements. Characteristic for a reflector were the elements 'attending to feelings', 'association', and/or 'integration'. A critical reflector had to show evidence of some or all of the elements of a reflector and in addition a change of perspective. The elements of reflection can be found again in the work of Findlay et al. (2010) together with a similar mapping strategy from the elements to the three levels of reflection. They call the elements of reflection 'deep analytic NRAT' and the levels 'broad classification NRAT'. Ip et al. (2012) focussed only on the three levels of reflection found in the model of Wong et al. (1995).

Hatton and Smith (1995, p. 48) distilled four types of writing, drawing on the theory of Schön (1983). A writing can be either descriptive/technical, descriptive reflective, dialogic reflective, or critical reflective. They argued that their data showed evidence that these types are indeed different types of reflection, each with its own features, use, and function. The first type is often seen as instrumental, as a starting point which can serve to engage with one of the types of reflection (Hatton and Smith, 1995, p. 46). The dialogic type is more 'exploratory', while the critical type is 'demanding critical' (Hatton and Smith, 1995, p. 46). In their study, they found that most of the coded units were descriptive, followed by dialogic reflective, and only a very few were critical reflective (Hatton and Smith, 1995, p. 41). They saw a potential reason for this in that the descriptive reflective type might be easier to master than either the dialogic or the reflective type of reflection (Hatton and Smith, 1995, p. 46).

The coding categories of Kember et al. (1999) are described as levels of reflection and not as types of reflection as in the work of Hatton and Smith (1995). The categories are ordered by their degree of reflectivity. Some of the categories are on the

same level, indicating an equal degree of reflectivity. One example is 'content reflection' and 'process reflection' (Kember et al., 1999, p 25), which are both on the same level. In order to differentiate the highest level of reflection, which they saw in 'premise reflection,' they introduced a second, lower level of reflection with the categories 'content', 'process,' and 'content and process' reflection. Other researchers have proposed other ways of arranging these levels, for example Wallman et al. (2008, p. 4) linearised the model of Kember et al. (1999) and put each category on its own separate level. At the highest level is 'premise reflection,' and at the lowest level is 'habitual action'. Later, Kember et al. (2008) remodelled the coding schema, giving up the idea of having categories on the same level, and instead of seven categories, the new model consists now of four categories: non-reflective, understanding, reflection, and critical reflection. The categories are again ordered by their level of reflection: from non-reflection to critical reflection.

As already mentioned, the breadth quality of reflection is foundational for the analysis of reflective texts. The next section will especially focus on this quality of reflection. Although the constituents of the models listed in Table 1 all seem very different, they do share similarities. The next section will distil these common categories. These categories will be later used for a coding schema with which a large corpus of writings will be annotated.

2.3 MODEL FOR REFLECTION DETECTION

Section 2.2 'Models to analyse written reflection' provided an overview with regard to the models used in the research to analyse written reflection using principles of the content analysis. The models outlined in Table 1 hint at the variety of possibilities to conceptualise reflection. The most general observation is that reflection is seen as a construct that comprises several categories. Each category represents a facet of

reflection. The categories are used as the coding schema for the content analysis of the text in order to analyse writing with regard to reflection.

Although these models vary, they share some commonalities. The following section describes the process used to evidence a set of categories that shares frequently found features of the outlined models of reflective writing. These categories are derived by analysing the breadth (descriptive) quality of the models. In the evaluation, each category serves as a test case for the automated detection of reflection.

The final set of categories are labelled with: 'Description of an experience', 'feelings', 'personal', 'critical stance', 'perspective', and 'outcome'. They form the categories of the theoretical model of reflection detection that guides this research process, and they are used as coding schema for the datasets to train and test the machine learning models. An earlier version of the categories was presented in Ullmann et al. (2012).

First, an outline of these six common categories of reflection is given (the full description of these categories can be found in Section 2.3.2 'Common reflection categories'). Subsequently, I show that these six categories can be found in many of the model constituents listed in Table 1. The derived common categories of reflection are:

Description of an experience: This category captures the subject matter of the reflective writing. It contains a description of what or how something occurred, recaptures important parts of the experience, provides the context, and/or the reason for the writing. For example, Schön (1987, p. 26) emphasised thinking about past experiences with the concept of reflection-on-action. Boud et al. (1985, chapter 1) saw reflection as thinking about past experiences.

Feelings: Reflection is not only seen as a cognitive process, but feelings are also often recognised as an important part of reflection. Boud et al. (1985, chapter 1) wrote about making use of helpful feelings and removing or containing obstructive ones.

Dewey (1933, p. 9) noted that a trigger for reflection can be the sensation of perplexity, hesitation, doubt, or surprise.

Personal: Reflection is often from a personal nature. This is about one's assumptions, beliefs, the development of a personal perspective, and the knowledge of self. Boyd and Fales (1983, p. 101) defined reflection as '(...) the process of creating and clarifying the meaning of experience (...) in terms of self (self in relation to self and self in relation to world)'.

Critical stance: Expressing an alert, critical mindset is an important part of reflective writing. A critical stance involves being aware of problems and being able to identify or diagnose such problems. This is about questioning opinions and assumptions, analysing and evaluating problems, judging situations, testing validity, drawing conclusions, and making decisions. Mezirow (1998, p. 186) wrote in the context of reflecting about assumptions that 'critical self-reflection of an assumption (...) involves critique of a premise upon which the learner has defined a problem'.

Perspective: The writer considers other perspectives. For example, the perspective of someone else, a theory, or the social, historical, ethical, moral, or political context. Moon (2004, p. 214) noted in her generic framework for reflective writing the importance of considering external ideas and multiple perspectives.

Outcome: The reflective writing contains a description of the lessons learned, better understanding of the situation or context, new insights, a change of perspective or behaviour, and awareness about one's way of thinking. It also contains statements of intention and planning for the future. Boud et al. (1985, p. 20) saw as the **outcomes** of reflection '(...) a personal synthesis, integration and appropriation of knowledge, the validation of personal knowledge, a new affective state, or the decision to engage in some further activity'. Mezirow (1991, chapter 4) wrote with regard to the outcome of the reflection that 'reflective learning can be either confirmative or transformative'. This means that either meaning schema can be confirmed, or a new meaning schema is

developed. Killion and Todnem (1991) emphasised reflection-for-action that comprises planning for future action based on the analysis of previous action.

Section 2.3.1 'Evidencing common categories of reflection' provides evidence for these six categories based on the literature of the analysis of reflective writings. Section 2.3.2 'Common reflection categories' presents a synopsis for each of the common categories of reflection, and Section 2.3.3 'Model critique' offers a critique of the model categories for reflection detection.

2.3.1 *Evidencing common categories of reflection*

This section provides supporting evidence for the claim that the categories outlined above are indeed frequently found in the investigated models of reflective writing.

The main source of evidence is the description of the models by their authors. They are the models outlined in Table 1 of Section 2.2 'Models to analyse written reflection'. Each of the models has been investigated regarding the six categories of reflection.

The following paragraphs combine the supporting evidence for each of the common reflection categories (description of an experience, feelings, personal, critical stance, perspective, and outcome) found in the models. Each category contains a description derived from the relevant parts of those models. In case the terminology used in one of the models does not directly refer to the category, the link between the model description and the category is outlined. In most cases, the relevant part of the model is directly mappable to the model. In some cases, this link cannot be established explicitly from the model, but it can be from the context described in its accompanying research. For such cases, the explaining text marks them as implicit. Cases where a model does not provide any evidence that justifies its categorisation are also marked clearly. Table 2 summarises this analysis. For each paper, the table indicates whether the model could be mapped to the common reflection categories.

Some of the models listed in Table 1 are very similar. To prevent that a model gets included in the synthesis several times, the following process was applied. If there are two models which are similar, then the model that pre-dates the other is maintained, whereas the model published later is not included in this synthesis.

The models that are similar, and thus not included in the following analysis, are the models of Hawkes and Romiszowski (2001) and Hawkes (2001, 2006), which are very similar to the model of Sparks-Langer et al. (1990). The models of Chirema (2007, p. 201) and Findlay et al. (2010, p. 86) (as well as the later paper of Findlay et al. (2011, p. 5)), which are very similar to the model used in Wong et al. (1995). The models of Bell et al. (2011, p. 801 f.) and Wallman et al. (2008, p. 9 f.) are very similar. Both are based on the model described in Kember et al. (1999). The model of Clarkeburn and Kettula (2011, p. 443) is based on the model of Kember et al. (2008, p. 379). The model of Hatton and Smith (1995) was used in the work of Pee et al. (2002, p. 578), Mena-Marcos et al. (2013, p. 151, 155), and Poom-Valickis and Mathews (2013). Mena-Marcos et al. (2013, p. 151, 155) used two models: Hatton and Smith (1995) and Tillema (2004). The first model is covered by the analysis of the model described by Pee et al. (2002, p. 578). The description of the latter model is based on the work of Mena-Marcos et al. (2013, p. 151,155).

Further, if a model in Table 1 does not show evidence of one of the categories because it only describes the reflection levels, or does not outline the coding categories, it is not listed. This concerns the following models: Sumsion and Fleet (1996, p. 125 f.), which referred in their work to the models of Surbeck et al. (1991) and Boud (1994). However, they do not provide information on their final coding schema. They only stated that they used the levels highly reflective, moderately reflective, and not reflective, but did not outline their model. Ip et al. (2012) referred to the level model of Wong et al. (1995), but they did not provide further information on the criteria that led to the categorisation of non-reflector, reflector, and critical reflector.

Description of an experience: The levels of the model of Sparks-Langer et al. (1990, p. 27) refer to the description of events and the description of explanations of the events. Both is related to 'description of an experience'. Wong et al. (1995, p. 57) referred in their marking schema to past experiences. They phrased it as relating to the old (see element association and integration). Further, they addressed the importance of 'prior knowledge or beliefs' (see elements 'integration' and 'validation'). The description of past experiences and prior knowledge relates to this category. McCollum (1997, p. 125, 127 f.) based her work on Tsangaridou and O'Sullivan (1994). This model explicitly mentioned the importance of the event description, and generally of what occurred. Content and process reflection as used in the model of Kember et al. (1999, p. 23) are both concerned with the description of the 'subject matter of the reflection'. The model of Fund et al. (2002, p. 492) contained the category called description. With this category, they captured the issues of a lesson, the way in which it was taught, and how the students described themselves. In the model of Hamann (2002, p. 5), no explicit references to a coding category related to description of an experience could be found. Pee et al. (2002, p. 578) based their model on Hatton and Smith (1995). Their model foresaw the category descriptive reflective, which relates to description of an experience as it contains the description of events and the provision of a rational for events or action. Williams et al. (2002, p. 15) highlighted the importance of the description of 'the learning event, issue, or situation'. The model of Boenink et al. (2004, p. 372) did not emphasise description of an experience, and thus, was not coded. The model of O'Connell and Dyment (2004) contained a schema based on the taxonomy of Bloom. A direct reference to description of an experience was not found. The model of Plack et al. (2005, p. 206 f.) had the element 'return to experience', which was used to code the experience descriptions. The first level of the model of Ballard (2006, p. 20 f., 29) - technical rationality - was used to hold the concerns of teachers on student behaviour.

Mansvelder-Longayroux (2006, p. 28, 58, 79 f.) and Mansvelder-Longayroux et al. (2007, p. 53 f.) described the thinking activity of recollection, which comprised the recall of past memories. The model of Plack et al. (2007, p. 287) dedicated a level to evidence 'data gathering' in the writings. Their explanation of this level contained direct evidence for this category, given that they wrote 'students describe the experience for the purpose of understanding' (Plack et al., 2007, p. 287). The coding reflection category for the model of Kember et al. (2008, p. 372 ff.) described the importance of relating concepts to personal experiences, which means that the writing should show evidence of an experience description, a concept, and the relationship between both. Wallman et al. (2008, p. 9) saw the 'description of the course of events' as part of their category 'habitual action'. They highlighted the importance of experience for reflection, given that they wrote that reflection is '(...) a situation identified in relation to actual experience' (Wallman et al., 2008, p. 9). The first category of the model of Chamoso and Cáceres (2009, p. 202, 205) is called 'description'. This entailed student descriptions with regard to their learning. The 'focus' dimension of the model of Lai and Calandra (2010, p. 434) gathered the descriptions of concerns regarding practice. The focus ranged from routine to transformative, capturing facets of the reflection 'focus'. The model of Fischer et al. (2011, p. 170) described that the non-reflective level is 'reporting only', whereas the level of 'reflection on experience' had to contain descriptions on the writer becoming aware of specific ways of seeing situations, or personal behaviour, feelings, and values. The first indicator of the model of Birney (2012, Appendix D) referred to the importance of describing the context to set the stage for writing. The category 'writing spectrum' on the level of introspection of the model of Wald et al. (2012, p. 48) referred to the elaborated description of the writer's impressions. The model used by Mena-Marcos et al. (2013, p. 151, (1a)) did not contain a reference to this category. The first category from Poldner et al. (2014, p. 358) is on the description of 'what,

when, and how regarding action', and of the context. The coding schema of Prilla and Renner (2014, p. 186) contained a category called 'Description of an experience and mentioning of an issue'.

Feelings: Sparks-Langer et al. (1990, p. 27) did not refer in their model to feelings, thus the cell in Table 2 is marked as absent. Wong et al. (1995, p. 57) explicitly referred to feelings. They emphasised the importance of making use of positive emotions and discarding negative feelings (code 1). The model used by McCollum (1997, p. 125) did not explicitly refer to feelings. However, (McCollum, 1997, p. 127) described reflection as containing 'thoughts and feelings' (McCollum, 1997, p. 127). In her analysis, McCollum captured the pupils' feelings (McCollum, 1997, p. 79). Kember et al. (1999, p. 21) wrote that 'introspection' concerns the 'feelings or thoughts about ourselves'. In addition 'content', 'process', and 'premise reflection' do recognise feelings as part of reflection (Kember et al., 1999, p. 23). Feelings were addressed by the model of Fund et al. (2002, p. 492). It captured the concerns that rely on feelings and intuition. The model of Hamann (2002, p. 5) mentioned attention to emotional needs. Pee et al. (2002, p. 578) did not explicitly mention feelings in their model. Feelings were addressed in three of the five criteria developed by Williams et al. (2002, p. 15). Boenink et al. (2004, p. 372) recognised doubt and uncertainty as important feelings for reflection. Feelings were not part of the model of O'Connell and Dyment (2004, p. 163). 'Attending to feelings' was one of the categories of the model of Plack et al. (2005, p. 206 f.). Ballard (2006, p. 20 f., 29) noted the feeling of being concerned about something. The model of Mansvelder-Longayroux (2006) and Mansvelder-Longayroux et al. (2007) did not contain a reference to feelings. Plack et al. (2007, p. 287) mentioned feelings on the levels of 'knowledge and comprehension' and 'analysis'. The model of Kember et al. (2008, p. 372 ff.) did not mention feelings. Several references to feelings can be found in the model of Wallman et al. (2008, p. 9). The level of introspection emphasised the relationship between

personal feelings and tasks. Content reflection is a description of what was felt, and process reflection is a description of how the function of feeling was performed. The category of feelings did not play a role in the model of Chamoso and Cáceres (2009, p. 202, 205). Lai and Calandra (2010, p. 434) referred to the feeling of concern, frustration, and excitement. Fischer et al. (2011, p. 170) addressed feelings on the low level of 'reflection on experience'. (Birney, 2012, p. 188, Appendix D) coded feelings as part of the indicator 'self-awareness'. The model of Wald et al. (2012, p. 48) dedicated an entire category to feelings, which they called 'attending to emotions'. Poldner et al. (2014, p. 358) did not refer to feelings in their model. The model of Prilla and Renner (2014, p. 186) directly addressed 'emotions'.

Personal: Sparks-Langer et al. (1990, p. 27) mentioned 'personal preferences' (level 4). Wong et al. (1995, p. 57) stressed in the description of the element 'appropriation' the role of 'making knowledge one's own', the 'sense of identity', and that 'New knowledge, feelings or attitudes becoming a significant force in own life'. The model used by McCollum (1997, p. 125) did not explicitly refer to something personal, but the teachers were asked to write their account from a personal perspective (McCollum, 1997, p. 127). The model of Kember et al. (1999, p. 23) referred to the categories 'content', 'process', and 'premise reflection', which concern one's perceptions, feelings, and actions. Their description of 'premise reflection' aimed at becoming aware of one's assumptions, values, or beliefs. These thinking frames are used to make sense of situations. They are specific perspectives that guide actions. 'Premise reflection' is the process of making such unconsciously held beliefs accessible to critical review. The personal perspective is important in the model of Fund et al. (2002, p. 492), which dedicated an entire dimension to it. According to the model of Hamann (2002, p. 5), a sign of a common-sense thinker is self-orientation, in addition to reliance on personal experience. An alert thinker roots his or her line of reasoning in the knowledge of self. Pee et al. (2002, p. 578) referred to a discourse with one's self

in the type called 'dialogic reflection'. The criteria mentioned in the model of Williams et al. (2002, p. 15) were centred on the personal experiences of the writer, their prior knowledge, feelings, or attitudes. Boenink et al. (2004, p. 372) referred to the 'personal perspective'. Personal reflections were frequently detected as a distinct type in the model of O'Connell and Dyment (2004, p. 163). Plack et al. (2005, p. 206 f.) explicitly mentioned the personal dimension of reflection. They looked with the category 'return to experience' for evidence of students describing personal relevant information. The writings assessed by Ballard (2006, p. 20 f., 29) were about the personal experiences of participants. However, the model did not mention explicitly a personal dimension. The model of Mansvelder-Longayroux (2006, p. 28, 58, 79 f.) and Mansvelder-Longayroux et al. (2007, p. 53 f.) implicitly addressed this category, given that many of the model subcategories refer explicitly to the writer perspective. Examples are the subcategories 'description of what you did or plan to do (and why)', 'description of how you approached something (...)', 'examining what you have learned', 'evaluating your knowledge (...)', etc. Plack et al. (2007, p. 287) emphasised that accounts should reveal the writer perspective. They were searching for evidence of explanations for '(...) what happened from his/her perspective' (Plack et al., 2007, p. 287). The model of Kember et al. (2008, p. 372 ff.) acknowledged the importance of personal experience for reflection. The level of introspection in the model of Wallman et al. (2008, p. 9) clearly referred to thoughts and feelings with regard to oneself. The examples that illustrate the categories of the model of Chamoso and Cáceres (2009, p. 202, 205) suggested the importance of personal involvement. The category 'argumentation' was illustrated with the following examples: 'To me it seems very important that (...)', 'In my opinion these classes are good because (...)', or 'I think it would be interesting and fun as well to learn different games in which geometry is present (...)' (Chamoso and Cáceres, 2009, 205). These examples indicate that personal statements were captured. Personal involvement is also important in the model of Lai

and Calandra (2010, p. 434). The model of Fischer et al. (2011, p. 170) had many references to the personal character of reflection. It referred to the 'specific perception, meaning or behaviour of one's own' (Fischer et al., 2011, p. 170), and the awareness of one's feelings, value judgements, and underlying beliefs. Birney (2012, Appendix D) identified self-awareness as one of the highest ranked reflection identifiers. Wald et al. (2012, p. 48) saw indicators of 'presence' as important for reflective writings. The highest level is reached if the text shows evidence that the writer is fully present. The examples provided by Mena-Marcos et al. (2013, p. 151, (1a)) to illustrate the model categories are written from a first-person perspective, which suggest that personal statements are important. Similarly, the examples for the categories evaluation, justification, dialogue, and transfer given by Poldner et al. (2014, p. 358) are indicative of the importance of personal messages. Several of the examples that describe the model categories of Prilla and Renner (2014, p. 186) contained self-references that indicate the importance of personal utterances in writings.

Critical stance: Through levels 4 to 7, Sparks-Langer et al. (1990, p. 27) emphasised the importance of explanation and providing a rationale. This indicates 'critical stance'. Wong et al. (1995, p. 57) used concepts such as re-assessment (element association), finding relationships (integration), and testing consistency (validation). This is indicative of having a 'critical stance'. The model used by McCollum (1997, p. 125) referred to evidence for justification and critique. Premise reflection is associated with the 'assessment of efficacy', and premise reflection requires a 'critical review of presuppositions' (Kember et al., 1999, p. 23). In the model of Fund et al. (2002, p. 492), critical stance is related to what is called the 'critical bridge'. Indicators of critical stance in the model of Hamann (2002, p. 5) were 'conclusions to be tentative', 'Means-end thinking; strategic thinking', and 'Reasoning grounded in knowledge of self'. The model of Pee et al. (2002, p. 578) referred to reasoning in three of the types of reflection. Analysis and verification are two criteria of the model of Williams et al.

(2002, p. 15). For Boenink et al. (2004, p. 372), evidence of a balanced, weighted approach that considers all perspectives is important, as well as the evaluation of one's position. The model for cognitive thinking used by O'Connell and Dyment (2004, p. 164) contained the elements analysis, synthesis, and evaluation, which can be linked to critical stance. The notion of critical stance is stated in several coding categories of the model of Plack et al. (2005, p. 206 f.). They acknowledge it in the description of strategies, critiquing, re-evaluation, and validation. Ballard (2006, p. 20 f., 29) recognised the importance of clarifying assumptions and addressing consequences and personal biases. Four of the six categories of the model of Mansvelder-Longayroux (2006) and Mansvelder-Longayroux et al. (2007) directly addressed critical stance. They are the categories of 'evaluation', 'analysis', 'critical processing', and 'diagnosis'. The category 'diagnosis', for example, contained subcategories that explore different facets of having difficulties or finding reasons for not achieving something. Showing evidence of critical stance is visible on all three levels of the model of Plack et al. (2007, p. 287). On the level of 'knowledge and comprehension', the critical stance is described as being aware of gaps in knowledge. The two other levels directly address analytical and evaluative stance. The model of Kember et al. (2008, p. 372 ff.) contained the category 'critical reflection'. Whereas the categories of 'habitual action' and 'understanding' are mostly uncritical, the category of 'critical reflection' asks for the evidence of a 'critical review of presuppositions' (Kember et al., 2008, p. 372 ff.). Evidence of critical stance can be found in the model of Wallman et al. (2008, p. 9). They stated that reflection is concerned with the analysis of problems. Further, someone who shows evidence of content reflection questions behaviour. Premise reflection contains the 'analysis of the whole situation/problem' (Wallman et al., 2008, p. 10). The 'argumentation' category of the model of Chamoso and Cáceres (2009, p. 202, 205) is about argumentation, justification, and conclusion drawing with regard to one's learning. The model of Lai and Calandra (2010, p. 434)

emphasised the importance of inquiry, which is that the text shows evidence of writers critically questioning their actions, situational context, and beliefs. Questions may arise from the awareness that something is not right, which may arise from the feeling of frustration, or unexpected outcomes of a situation. The model of Fischer et al. (2011, p. 170) highlighted the importance of assessment, judgement, questioning, and awareness of underlying assumptions for reflection. Assessment is related to efficacy assessment. Also important is awareness of the influence of values on judgements, and that judgements might be uninformed. Being able to question the adequacy of one's understanding and underlying beliefs and awareness of assumptions made while judging a situation are seen as 'reflection on awareness'. The model of Birney (2012, Appendix D) contained several indicators related to critical stance. These are the indicators of 'analysis', 'synthesis', correct identification of issues, and the discussion of 'implications of actions'. Issue identification is important because, first, one has to become aware that there is an issue and that it is important, before it can be analysed. Acknowledging that there are issues requires critical stance. The model of Wald et al. (2012, p. 48) contained an entire category for the description of a conflict, dilemma, challenge, or issue of concern. The category foresaw four levels that range from uncritical writings (the writer did not describe a problem or was not aware of it) to a full account of the problem considering several perspectives, and the exploration of alternative interpretations. Furthermore, the model contained a category dedicated to 'analysis'. The model of (Mena-Marcos et al., 2013, p. 151, (1a)) referred to evidencing rule usage with premise and conclusion structure – an analytical element. In addition, they mentioned positive or negative 'appraisal' – an evaluative element. The model of Poldner et al. (2014, p. 358) reviewed the category of critical stance in two ways. Their model contained the categories 'evaluation' and 'justification'. References to 'critical stance' can be found in the model of Prilla and Renner (2014, p. 186) in the categories of 'mentioning an issue', which can be seen as

an awareness of problems; in the category 'interpreting or explaining behavior', especially because they include evidences of giving a rationale for behaviour; in the category 'responding to explanations of an experience by challenging or supporting assumptions'; and in the category 'responding to the explanation of an experience by providing reasons for the issue'.

Perspective: This category regards the consideration or integration of alternative perspectives in one's thinking. Several models mention perspective transformation, a fundamental change of the belief system. This aspect of reflection is usually discussed in the context of the aims of reflection, and it is seen as an outcome of reflection. Therefore, it is discussed in the category outcome. The models often mention the importance of the personal perspective (one's beliefs and assumptions). A personal perspective is seen as part of the category 'personal'.

Three types of perspectives are mentioned in the model of Sparks-Langer et al. (1990, p. 27) – considering a theory (level 5), considering context factors (level 6), and explaining something with respect to the ethical, moral, or political perspective (level 7). The explanations of the model of Wong et al. (1995, p. 57) were not sufficiently conclusive to sort this model into this category, and thus the category perspective is marked as absent in Table 2. The focus dimension of the model used by McCollum (1997, p. 125) considered the perspective of the situational context and the (sensitizing) perspectives of the '(...) social, moral, political, or ethical issues (...)' (McCollum, 1997, p. 128). The model of Kember et al. (1999, p. 23) does not address this category explicitly. They mentioned 'premise reflection' as the highest form of reflection as it might lead to perspective transformation. Premise reflection is the awareness about one's beliefs or assumptions. Perspective transformation is seen as part of the 'outcome' category and the awareness of one's assumptions as part of the category 'personal'. Fund et al. (2002, p. 491) wrote with regard to the category 'critical bridging' that it contained a 'discussion of possible alternative opinions',

which, in essence, is the act of considering another perspective. In addition, they emphasised the consideration of the context of a lesson in order to analyse the behaviour of a person (Fund et al., 2002, p. 492). Here, the perspective to be considered is the context. Several perspectives are mentioned in the model of Hamann (2002, p. 5). The model categories for 'alert thinker' mentioned, for example, attention to social, ethical, or political perspectives, awareness of its own position as a role model, and the social and political context of a classroom. The 'critical' type of reflection in the model of Pee et al. (2002, p. 578) referred to the socio-political as an important perspective to consider when making decisions. In addition, the exploration of alternatives was mentioned. In the model of Williams et al. (2002, p. 15), no direct references could be found, and thus, it was not recorded. Considering perspectives is an important factor in the model of Boenink et al. (2004, p. 372). The more perspectives are considered and weighted against each other, the higher is the scoring. The model of O'Connell and Dyment (2004, p. 164) had no reference to this category. The importance of seeking multiple perspectives, and exploring experience from several perspectives, is evident in the model of Plack et al. (2005, p. 206 f.). The model of Ballard (2006, p. 20 f., 29) contained no direct reference to this category. The category 'critical processing' of the model of Mansvelder-Longayroux (2006) and Mansvelder-Longayroux et al. (2007) is defined as the comparison between one's opinions and those of others in order to compare them according to credibility. This category describes the comparison of perspectives in order to make informed decisions. Plack et al. (2007, p. 287) recognised multiple perspectives as part of reflection. They wrote that the 'more skilful reflector would analyse experience from several different perspectives beyond the self' (Plack et al., 2007, p. 287). Kember et al. (2008, p. 372 ff.) provided two examples of perspectives that a writer could consider when reflecting. First, a concept can be interpreted with perspective to a personal experience. Second, situations can be interpreted against taught knowledge. Wallman

et al. (2008, p. 10) wrote on the importance of considering alternative methods because they may challenge prejudices. A reinterpretation of experience can be useful for seeing problems differently, which may change actions. Chamoso and Cáceres (2009, p. 199) emphasised in the theoretical part of the paper that, for reflection, it is important to be aware of other perspectives. However, in their model, they did not particularly emphasise the importance of considering other perspectives. Lai and Calandra (2010, p. 434) indicated openness to the considerations of others and the active search of the perspectives of others as important elements of their 'inquiry' dimension. The model of Fischer et al. (2011, p. 170) addressed awareness of the personal perspective of perceiving and judging situations. This is seen as part of the category 'personal', but not of this category that refers to perspectives beyond the self. The model of Birney (2012, Appendix D) explicitly addressed the importance of the examination of 'multiple perspectives'. Further, the indicator 'links are made to the broader social structures' considered the perspectives of the 'historical context', 'social context', 'ethical context', and 'legal context'. Discussing a topic through the lens of specific contexts adds perspective to the narrative. The highest level of reflection for the criterion 'description of conflict or disorienting dilemma' of the model of Wald et al. (2012, p. 48) contained references to perspectives. They wrote that for a full description of the dilemma, the text has to show evidence of '(...) multiple perspectives, exploring alternative explanations (...)' (Wald et al., 2012, p. 48). The model of Mena-Marcos et al. (2013, p. 151, (1a)) did not address this category. Perspectives were not mentioned in the model of Poldner et al. (2014, p. 358). However, the 'dialogue' category of the model of Poldner et al. (2014, p. 358) could indicate this direction. It contained the subcategory 'dialogue with student theory', which could be seen using theory as the context or guiding perspective of an inner dialogue on the classroom situation. The 'linking' categories of the model of Prilla

and Renner (2014, p. 186) can be seen as considering other perspectives. It considers linking own experience with experiences of others.

Outcome: Sparks-Langer et al. (1990, p. 27) did not particularly mention outcome as a category. The outcome category is not recorded for this model. Wong et al. (1995, p. 57) explicitly labelled one of their model's elements as outcome. The outcome of a reflection can be a transformation of perspectives, a change in behaviour, a readiness for something, and/or a commitment to action. The element of 'appropriation' can also be seen as an outcome. It describes the process of learning, where new knowledge becomes part of the identity. An outcome of reflection was not explicitly mentioned by McCollum (1997). The outcome of a 'premise reflection' in the model of Kember et al. (1999, p. 23 f.) was a 'significant change of perspective'. 'Process' and 'premise reflection' can help to become better aware of one's beliefs, thinking, and feelings, and how these influence action. Fund et al. (2002, p. 491) did not explicitly mention an outcome category of reflection. Implicitly, they recorded 'personal insights', and the reaching of conclusions. All of these can be seen as outcome of reflection. The models of Hamann (2002, p. 5) and Pee et al. (2002, p. 578) did not address a distinct outcome category. An outcome of reflection can be a new understanding, a clarification of the issue, learning of a skill, the resolution of a problem, or a informed judgement on how something will influence future behaviour (Williams et al., 2002, p. 15). The model of Boenink et al. (2004, p. 372) did not highlight a specific outcome category. The model aimed at learning the importance of multiple perspectives and managing situations of uncertainty. The thinking taxonomy used by O'Connell and Dyment (2004, p. 164) can be understood as evidence of learning in the categories of knowledge, comprehension, application, analysis, synthesis, and evaluation. Outcomes in the model of Plack et al. (2005, p. 206 f.) are better understanding, planning for the future, and the appropriation of new meaning into the one's way of being. An outcome dimension was not mentioned in the model

of Ballard (2006, p. 20 f., 29). The model of Mansvelder-Longayroux (2006) (and Mansvelder-Longayroux et al. (2007)) contained several references to an outcome of reflection. An outcome can be an examination of the progress made thus far, what was learned from the experience, and the consequences associated with actions taken. On the level of 'synthesis and evaluation' of the model of Plack et al. (2007, p. 287), an outcome of reflection can either be that something is learned, certain conclusions are drawn, or future strategies are planned. An outcome of a reflection according to Kember et al. (2008, p. 372 ff.) can be awareness regarding one's beliefs/assumptions, or the change of a belief system. They wrote that evidence of critical reflection is '(...) a change in perspective over a fundamental belief' (Kember et al., 2008, p. 375). The awareness of, and ability to, question prejudices, and the ability to reinterpret and better understand a situation in order to act differently next time a similar situation occurs, are outcomes described in the model of Wallman et al. (2008, p. 10). The 'contribution' category of the model of Chamoso and Cáceres (2009, p. 202, 205) listed as evidence of an outcome that students are involved, provide suggestions, and actively contribute improvements. The dimension of 'change' of the model of Lai and Calandra (2010, p. 434) listed several outcomes of reflection. Outcomes are the development of new insights that improve practice, and a 'reframing of perspective leading to fundamental change of practice' (Lai and Calandra, 2010, p. 434). Outcomes of reflection according to the model of Fischer et al. (2011, p. 170) was a heightened awareness of the reasons for one's way of perceiving and judging. Outcomes of the model of Birney (2012, Appendix D) can be seen in the evidence of 'learning', 'understanding', 'changes in beliefs', and intentions about 'future practices'. According to Wald et al. (2012, p. 48), the outcome of a reflection can be either confirmatory or transformative. The generation of an artefact to solve a problem can be seen as an outcome in the model of Mena-Marcos et al. (2013, p. 151, (1a)). The 'transfer' category of Poldner et al. (2014, p. 363) described an outcome dimension,

given that it categorises statements on what was learned and what to do in the future. Categories of the model of Prilla and Renner (2014, p. 186) related to outcome are the provision of 'solution proposals', 'learning', the intention to change, or the evidence of 'implementing change'.

The following Table 1 provides a summary of the supporting evidence found in the models of reflection for the common categories of reflection. The table lists all discussed models along with a mapping of their model constituents onto the six common reflection categories. Thus, the table provides evidence of the degree to which each model can be mapped to the common reflection categories.

Table 1 is read as follows. Each row represents a reflection model. Each cell indicates whether evidence was found in the description of the model that relates to one of the common reflection categories. The meaning of the symbols used in each cell is ✔ evidence of the presence of the category found in the model. ✓ Evidence implicit (see explanation on p. 30). ✗ Evidence not present. The column headings represent the common categories. The same table, but with verbal descriptions of the cell contents, can be found in Appendix B 'MAPPING OF MODELS OF REFLECTION TO COMMON CATEGORIES OF REFLECTION'.

Author(s)	Description of an experience	Feelings	Personal	Critical stance	Perspective	Outcome
Sparks-Langer et al. (1990)	✔	✗	✔	✔	✔	✗
Wong et al. (1995)	✔	✔	✔	✔	✗	✔
McCollum (1997)	✔	✓	✓	✔	✔	✗
Kember et al. (1999)	✔	✔	✔	✔	✗	✔
Fund et al. (2002)	✔	✔	✔	✔	✔	✓
Hamann (2002)	✗	✔	✔	✔	✗	✗
Pee et al. (2002)	✔	✗	✔	✔	✔	✗
Williams et al. (2002)	✔	✔	✔	✔	✗	✔
Boenink et al. (2004)	✗	✔	✔	✔	✔	✗
O'Connell and Dyment (2004)	✗	✗	✔	✔	✗	✔
Plack et al. (2005)	✔	✔	✔	✔	✔	✔
Ballard (2006)	✔	✔	✓	✔	✗	✗
Mansvelder-Longayroux (2006); Mansvelder-Longayroux et al. (2007)	✔	✗	✓	✔	✔	✔
Plack et al. (2007)	✔	✔	✔	✔	✔	✔

Continued on next page

Table 2 Continued from previous page

Author(s)	Description of an experience	Feelings	Personal	Critical stance	Perspective	Outcome
Kember et al. (2008)	✔	✗	✔	✔	✔	✔
Wallman et al. (2008)	✔	✔	✔	✔	✔	✔
Chamoso and Cáceres (2009)	✔	✗	✓	✔	✗	✔
Lai and Calandra (2010)	✔	✔	✔	✔	✔	✔
Fischer et al. (2011)	✔	✔	✔	✔	✗	✔
Birney (2012)	✔	✔	✔	✔	✔	✔
Wald et al. (2012)	✔	✔	✔	✔	✔	✔
Mena-Marcos et al. (2013)	✗	✗	✓	✔	✗	✔
Poldner et al. (2014)	✔	✗	✓	✔	✗	✔
Prilla and Renner (2014)	✔	✔	✓	✔	✔	✔

Table 2: Overview of the mapping of models to the common categories of reflection

Legend: ✔ Present. ✓ Implicit. ✗ Not present

Table 2 shows that the six common reflection categories are helpful to subsume the constituents of the models of reflective writings (Table 1). All models contain references to several of the categories 'description of an experience', 'feelings', 'personal', 'critical stance', 'perspective', and 'outcome'. Approximately one third of the models showed evidence of all common reflection categories.

All models mention elements that highlight the importance of **critical stance** or a critical mindset expressed in reflective writing. This category also subsumes a variety of critical thinking skills, e.g., the identification of problems, questioning, analysis/reasoning, evaluation, judging, and decision-making.

Almost all models search for evidence of **description of an experience** in a writing. The description of an experience can be seen as the introduction to reflective writing that opens the discussion space for the reflection. Level models usually assert that if a given writing is only descriptive (with no description of experience) – for example, it only states course events, it is rated as non-reflective. If a given writing shows evidence of a description of experiences only, it is usually classified on the lower levels of reflection. Evidence of a description of experience along with other common reflection categories leads, generally, to an assignment of the writings to higher levels

of reflection (see Section 3.1.3 'Relationship between the descriptive and level reflection quality').

Many models mention **personal** stance as important for a reflective writing. Many models refer directly to this category, but also, a large proportion of models implicitly consider this category when classifying reflective writings. Explicitly, they referred to this category as, for example, awareness of one's beliefs, values, assumptions, 'sense of identity', 'knowledge of self', 'personal perspective', 'self-awareness', or 'presence' of the writer. In addition to explicitly referring to the category 'personal', many models refer to this category in the way they phrased the model descriptions. For example, Plack et al. (2005, p. 207) phrased all three 'stage dependent' categories from the perspective of the student. They explained the category 'return to experience' with 'The student describes an experience replaying what he or she considers significant', and 'The student (...) begins to work with feelings (...)' (category ATTEND), or 'The student reappraises the current situation (...)' (category RE-EVAL) (Plack et al., 2005, p. 207). Mansvelder-Longayroux (2006, p. 28, 58, 79 f.) and Mansvelder-Longayroux et al. (2007, p. 53 f.) used the following phrasing: 'description of what you did or plan to do (and why)', 'description of how you approached something (...)', 'examining what you have learned', and 'evaluating your knowledge (...)'. Instead of having a separate category for capturing the personal stance in a given writing, the person perspective is encapsulated in the category description. In essence, this means that the process of assigning a category to a text unit consists of two steps: one to determine whether the unit is with regard to something personal, an another that involves deciding whether the unit fits the category.

Outcome of a reflection is evident in many models, especially in more recent models. Several types of outcome were of interest to the model authors. Some examples are significant change of perspective or behaviour, intention to do something, plan for

the future, learning, awareness of one's way of decision-making, understanding of the situation/context, insight, or the confirmation of the existing way of thinking.

The category of **perspective** is part of many models, which justifies its own category. In particular, this is about the evidence of a writer being aware of, and considering, alternative perspectives. The models mentioned, for example, the perspective of others, context factors, interpretation through the lens of a theory, social, historical, ethical, moral, or political perspectives, etc.

Feelings is addressed by 15 out of 24 models (63%). Many models recognise the importance of emotions for reflection. The models address feelings in general, both positive and negative, and particularly, feelings of concern, doubt, uncertainty, frustration, and excitement.

From this extensive analysis of reflection models, it can be concluded that there are, indeed, commonalities between such widely varying models. The six common reflection categories are useful to capture frequently mentioned constituents of the outlined models. These common categories serve in the following sections as the model categories for the automated detection of reflection. The next section describes each of these six categories.

2.3.2 *Common reflection categories*

Section 2.2 'Models to analyse written reflection' showed the variety of existing model descriptions to analyse reflective texts. In Section 2.3.1 'Evidencing common categories of reflection', the model constitutes were analysed with regard to their common features. Six themes occur frequently in most models, which are the description of an experience, addressing of feelings, awareness of one's personal perspective, being critical, considering other perspectives, and a description of

outcomes. These are common categories of many reflective writing models, as shown in Table 2.

Each of the common categories represents a test case for automated detection. In addition to these test cases, machine learning is tested as to the degree to which it can discern between reflective and descriptive text segments. This distinction is the common denominator of the models that analyse the depth of reflection (see the discussion on the quality of depth in Section 2.2 'Models to analyse written reflection').

The following boxes describe each common category of reflective writing with a short summary.

Description of an experience: This category captures the subject matter of the reflective writing. It contains a description of what or how something occurred, recaptures important parts of the experience, provides the context, and/or the reason for the writing.

Feelings: Reflection is not only seen as a cognitive process, but feelings are also often recognised as an important part of reflection. Often, the feeling of being concerned, having doubts, feeling uncertain about something, or frustration are reasons for a reflective thought process. However, feelings such as surprise or excitement are also mentioned. Whereas feelings can be a starting point for reflection, they can also be the subject of reflection. These are reflections on the influence of feelings on thinking and action.

Personal: Reflection is often from a personal nature. This is about one's assumptions, beliefs, the development of a personal perspective, and the knowledge of self.

Critical stance: Expressing an alert, critical mindset is an important part of reflective writing. A critical stance involves being aware of problems and being able to identify or diagnose such problems. This is about questioning opinions and assumptions, analysing and evaluating problems, judging situations, testing validity, drawing conclusions, and making decisions.

Perspective: The writer considers other perspectives. For example, the perspective of someone else, theory, the social, historical, ethical, moral, or political context.

Outcome: Several types of outcomes have been described. Retrospective outcomes were: Descriptions of the lessons learned, better understanding of the situation or context, new insights, a change of perspective or behaviour, and awareness about one's way of thinking. Prospective outcomes were: An intention to do something, and planning for the future.

The next section outlines several considerations, which are important for the understanding of these six categories.

2.3.3 *Model critique*

These six common categories subsume a large array of model constituents of existing models that are used to analyse reflective writing. They provide a common vocabulary that allows the discussion of important aspects of reflective writings.

Such categories are more general than the specific constituents of the individual models, while being sufficiently specific to capture important dimensions of reflective writings. They are more general because they have been created by subsuming several

similar aspects into one category. In addition, they are specific to the context of reflective writing because the categories capture important dimensions of reflective writing.

The model was created with the aim of inferring from the models of reflection a set of categories that represent important facets of reflective writing. The benefit of this approach is seen in that the machine learning algorithms are tested on several important categories common to many models. In addition, this research is not associated with a specific model, and therefore, testing the applicability of the automated detection more generally is possible. The limitation is that the model does not capture all the facets of individual models. The decision falls on the more general approach in order to derive more general insight on automated reflection detection instead of researching the characteristics of individual models.

As outlined in Section 2.2 'Models to analyse written reflection' the common categories represent the breadth (descriptive) dimension of reflective writings. These are categories that describe the writings without implying a hierarchy of reflection, as the depth (level) models do. The reason for this focus on breadth dimensions is that the descriptive categories found in texts have been used to infer the level of reflection via a mapping mechanism (see Section 3.1.3 'Relationship between the descriptive and level reflection quality'). Once the descriptive categories are found inside the text the mapping mechanism assigns a level of reflection to the entire text. This suggested to focus the study of the automated detection of reflection on the descriptive categories of reflection, as they first have to be determined before the level can be inferred by the mapping strategy. Besides the mapping of descriptive categories to levels of reflection, researchers also inferred the levels directly (see for example Kember et al. (1999) described in Section 2.2 'Models to analyse written reflection' on page 26). The depth dimension is captured in the model for reflection detection in the

common denominator of all level models that is the distinction between descriptive/non-reflective and reflective.

After this general critique of the model it follows clarifications regarding some of the common categories.

With regard to the categories 'critical stance' and 'perspective'. Considering other perspectives can be seen as a critical stance, because alternatives are compared and evaluated. In this context, the category perspective might belong to the category 'critical stance'. Here, the category critical stance is used to express analytical thinking based on one's reasoning. 'Perspective' can be contrasted with 'critical stance' insofar as it emphasises awareness of other perspectives and the line of thought stemming from outside the self. Further, several models mention perspective transformation. The transformation of perspective is an aim of reflection and was therefore subsumed into the category 'outcome'. The models also noted the importance of the personal perspective. The description of a personal perspective is part of the category 'personal' and not 'perspective' as it is about the personal standpoint, one's beliefs, assumptions and not the consideration of other perspectives.

The category 'feelings' can be seen as part of the category 'description of an experience' because a feeling can be an experience that triggers reflective writing. In the older version of this model (see Ullmann et al. (2012, p. 103 f.)), 'feelings' is part of the category 'description of an experience'. Depending on the context and the research aims, one might want to determine whether feelings belongs in its own category, or should be part of other categories.

2.4 SUMMARY

This section provided an overview on reflection, particularly on models used to analyse written reflections. Reflection definitions based on high impact papers were

used to discern what is meant by reflection in daily language and what it means in research. Section 2.2 'Models to analyse written reflection' illustrated the variety of the existing ways of modelling reflection in the context of the analysis of reflective writing. To structure this variety, two qualities of reflection, breadth and depth, were introduced. The relationship between both was depicted, and it was concluded that in many instances, reflection levels can be derived from its descriptive (breadth) constituents. This led to the decision to focus on the descriptive reflection elements. In Section 2.3.1 'Evidencing common categories of reflection', each of the descriptive elements of the models listed in Table 1 were classified into the themes 'description of an experience', 'feelings', 'personal', 'critical stance', 'perspective', and 'outcome'. A case was made for all models and common categories to be seen as useful to subsume many of the varying categories used by the models. These six categories are sufficient general to provide a general framework for reflection categories, and they are sufficiently specific to capture the most important categories of reflective writing. A summary of the six common categories and a critique concluded this chapter.

The outcome of this chapter is the synthesis of the six common reflection categories. They are used as the model categories for reflection detection.

3

RELATED METHODS AND BENCHMARKS

The previous chapter Chapter 2 'THE CONCEPT OF REFLECTION: THEORY AND MODEL' defined reflection and provided an overview on the models used to analyse written reflection. The current chapter focusses on relevant methods.

It contains two major strands. The first strand investigates the manual methods to analyse reflection in writings. The focus is particularly on methods that apply content analysis in order to gain insight on reflection from writings. The method for manual content analysis on reflection is close to what this thesis attempts to automate. The parallel is that both the manual content analysis and machine learning classifiers label text units according to reflection categories.

The second strand is on automated methods that have been applied to the concepts related to reflection. These are methods that automatically derive insight from written text on concepts such as critical thinking, argumentation, or epistemic activity. As outlined in the motivation of this thesis (see Chapter 1 'INTRODUCTION'), insufficient research exists on automated methods to analyse reflection in texts. Therefore, this literature review on methods extends its scope to automated methods, which are used to detect concepts related to reflection.

Both the theoretical part of this thesis and this chapter inform the methodology and research design of this thesis (see Chapter 4 'METHODOLOGY AND RESEARCH DESIGN').

Section 3.1 'Manual methods to detect reflection' starts with an introduction to content analysis, especially on the process for inferring information from text. This

relates directly to Section 3.1.2 'Relationship between analysis units and reflection categories', which reports on the techniques for evidencing reflection, and Section 3.1.4 'Manual reflection detection performance', which reports on the metrics that indicate the overall quality of this inference process. In Section 2.2 'Models to analyse written reflection', the descriptive (breadth) reflection dimension and the level or depth reflection quality have been introduced. There, the argument is made that often the level of reflective writing can be inferred by its descriptive (breadth) quality. This led to the decision of focusing this research on the descriptive elements of the models of reflective writing. Section 3.1.3 'Relationship between the descriptive and level reflection quality' delivers evidence of this connection. The most important part of this strand is Section 3.1.4 'Manual reflection detection performance'. In order to determine performance indicators for the automated methods to detect reflection, this section gathers evidence of the performance of raters that analyse reflective texts using the content analysis method.

The sections on the second strand of related automated methods (Section 3.2 'Related automated methods') start with dictionary-based (see Section 3.2.1) and rule-based approaches (see Section 3.2.2), and ends in a section on machine learning approaches (see Section 3.2.3). This order was chosen to contrast machine learning with the first two approaches. The discussion on these three approaches leads to an argument that motivates the choice of machine learning methods to detect reflection (see Section 4.1 'General methodological considerations'). The core part of this strand is in Section 3.2.3 'Machine learning approaches'. This section presents research that uses machine learning to predict model categories that are important for educational research. This section informs the selection of machine learning algorithms and provides insight on the performance of these methods in areas related to reflection.

3.1 MANUAL METHODS TO DETECT REFLECTION

Several approaches exist to measure reflection evidence. Such methods range from qualitative to quantitative tools. The tools to identify evidence of reflection are questionnaires (Kember et al., 2000; Sobral, 2000; Woerkom et al., 2002; Mamede and Schmidt, 2004; Sobral, 2005; Peltier et al., 2005; Woerkom and Croon, 2008; Lethbridge et al., 2011; Bogo et al., 2011; Andersen et al., 2014), interviews (Boyd and Fales, 1983; King and Kitchener, 1994; Corlett, 2013), structured worksheets, vignettes or cue questions (Pee et al., 2002; Boenink et al., 2004; Ip et al., 2012), tests (King and Kitchener, 1994, p. 116ff.), linguistic analysis (Shaheed and Dong, 2006; Luk, 2008; Reidsema and Mort, 2009; Forbes, 2011; Ryan, 2011; Birney, 2012; Ryan, 2012; Wharton, 2012; Ryan, 2014), and manual content analysis of reflective writings.

From this variety of available methods, the method content analysis is chosen as the most relevant method for this thesis. Questionnaires attempt to elicit qualities of reflective thinking from participants. As such, they do not directly contribute to the problem of analysing text. Interviews, structured worksheets, and vignettes or cue questions are methods that aim to trigger reflective thinking in the participants based on specific prompts. Analysis of the participant responses are subject to the method of content analysis. An emerging area of researching reflective writing can be identified in the field of linguistics. This is especially visible in the research of Birney (2012), which mapped linguistic indicators to subcategories of a comprehensive model of reflection. The focus of this thesis is on the more established method of content analysis to analyse written reflection.

As outlined above, the reason for this focus is that manual content coding as performed with the method of content analysis is close to the aim of this thesis because it closely resembles the problem space of the automated analysis of reflection.

Using this method, text units are annotated with labels. This is similar to the aim of this thesis because it strives to automatically label text units.

3.1.1 *Content analysis of reflective writings*

Content analysis has a long tradition. It can be traced to the systematic analysis of text during the 17th century (Krippendorff, 2012, p. 10). There are many forms (Hsieh and Shannon, 2005) and definitions of content analysis. One of the more prominent definitions is the following:

'Content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use' (Krippendorff, 2012, p. 24).

The notion of inference is one of importance. Krippendorff (2012, p. 42) indicates that content analysis relies mostly on abductive reasoning – sometimes called the 'inference to the Best Explanation'[1] – and less on deductive or inductive reasoning. '(...) the whole enterprise of content analysis may well be regarded as an argument in support of an analyst's abductive claim' (Krippendorff, 2012, p. 43).

Applying this to the problem of automated reflection detection means that the claim drawn from the text has to be backed by a form of justification. For example, if a unit of analysis is automatically annotated with a category of reflective writing (the claim), a form of justification has to back this inference. In the case of manual content analysis, a supporting argument for this claim can be that the majority of raters agree on the classification of this unit. A common measurement that abstracts from the individual case to the performance of the entire coding process is inter-rater reliability. In the case of automated methods, the performance of the classifier can be compared with the ground truth. Such ground truth can be a collection of units that are correctly

1 http://plato.stanford.edu/entries/abduction/

classified with high certainty. A comparison of the results of the classifier and the ground truth indicates the quality of the classifier. Section 4.2 'Evaluation criteria and metrics' summarises several of these evaluation metrics.

The next sections are dedicated to important characteristics of the manual content analysis of reflection.

3.1.2 *Relationship between analysis units and reflection categories*

As outlined in Section 2.2 'Models to analyse written reflection', the models for assessing reflective writings consists of several categories. During the coding process, text units are assigned to these categories. These text units are evidence for this category. However, how much evidence is required to support the claim that the text indeed exhibits this category? Is a single appearance already sufficient, or does it require further corroboration (substantiation) in order to count as evidence?

Substantiation describes the extent to which an element of reflection has to be elaborated until it is determined as a valid element. For example, a given text can mention an emotion, which would fit the category of 'feelings'. Would mentioning be a satisfying criterion for its classification as feelings?

To find answers to this question, the literature on the manual content analysis of reflection was consulted (see Section 2.2 'Models to analyse written reflection').

In many cases, the units are used to calculate aggregated statistics, for example, frequencies, percentages, or mean values (for examples, see Poldner et al. (2014); Mena-Marcos et al. (2013); Clarkeburn and Kettula (2011)). This entails that categories are assigned to all units of analysis and then aggregated either at the text or student level. It also means that every unit counts as evidence for a category.

Plack et al. (2005, p. 203) reported that if a single piece of evidence for a category of reflection is found, it is not required by the raters to search for any further evidence

information. In this case, it is not necessary to find all pieces of evidence of a certain category in the entire text. The consequence for the automated detection of reflection is that once the classifier determines a single unit of analysis, its work can stop.

Plack et al. (2005, p. 204) then examined this strategy and noted that two of three raters coded a category if it contained further substantiation, whereas one rater coded a category if any textual indicator was given. Plack et al. (2005, p. 204) saw the approach taken by the two raters as more adequate. This approach indicates that for the presence of a category, at least two evidences have to be found.

Wong et al. (1995) stated that if an argument is made more than once, it is coded only once. Coding is based on the criteria of substantiation (only if further evidence is given, is it coded) and textual evidence (interpretation is not accepted). Quotes from literature count as substantiation if they are not mere textbook knowledge. This hints at a minimum of two evidences to justify a classification. Further, they noted certain exceptions, such as cited literature or similar arguments.

Lai and Calandra (2010) recorded only the highest level of reflection. This means that once the highest level is found, no other evidence has to be considered. It also means that only one unit of analysis is necessary to determine the level of the entire text.

From the cases outlined above, it can be seen that there are many different proposals. The common approach treats each labelled unit as evidence for the presence of a category.

3.1.3 *Relationship between the descriptive and level reflection quality*

As outlined in the theoretical part of the thesis (see Section 2.2 'Models to analyse written reflection'), models can be characterised as breadth models that describe reflection categories, and depth models as those that describe reflection levels.

Research that records both the breadth and depth dimensions often describe a mapping technique between both qualities.

Several articles described a two-stage system of coding. In the first stage, the reflective texts are coded according to reflection breadth categories, and in a second stage, they are mapped to reflection levels. The latter are often based on the distinction between non-reflective, reflection, and critical reflection (Manen, 1977; Mezirow, 1990a).

The six breadth reflection categories of Wong et al. (1995) were mapped to the depth model of Mezirow (1990a). A non-reflector shows no evidence of any of the six categories. A person indicated as a reflector provides evidence for the categories 'attending to feelings' to 'appropriation', whereas a critical reflector shows all features of a reflector plus evidence of changes in perspective. This is a direct mapping of reflection breadth categories to a reflection depth model.

Sumsion and Fleet (1996, p. 126) used one-to-one mapping from reflection categories to reflection levels. The three categories reaction, elaboration, and contemplation, based on the model of Surbeck et al. (1991), are mapped to the three levels of not reflective, moderately reflective, and highly reflective.

Hamann (2002, p. 10 ff.) mapped several categories to four types of reflective thinking. Ten categories defined non-reflective thinking (see Table 1 for the description of the categories); reflective cognitive was defined with four categories; reflective affective, social, psychomotor with eleven categories; and reflective inter and intra-personal were defined with one category.

Abou Baker El-Dib (2007, p. 34) described a model were categories are mapped to four levels of reflection. Mapping follows an additive schema. For example, the category 'problem statement' consists of the following rules: if a problem statement is evident, but there is no reasoning, it belongs to the lowest level (level 0). A problem statement plus evidence of a single reason signifies level 1. Level 2 shows evidence of

a problem statement plus several reasons. Finally, level 3 shows all the characteristics of level 2 plus evidence that considers the larger context.

Chirema (2007) used the following mapping schema. A non-reflector shows no evidence of any category. A reflector shows evidence of one or more of the following categories: attending to feelings, association, and integration. A critical reflector has to satisfy the following categories: validation, appropriation, and/or outcome of reflection.

Similarly, Findlay et al. (2010, 2011) showed mapping between the deep analytic NRAT and the broad analytic NRAT. If none of the categories are present, the writer is classified as a non-reflector. A reflector shows evidence of attending to feelings, association, and integration. A critical reflection has to demonstrate validation, appropriation, and outcome.

Bell et al. (2011) mapping schema was based on Kember et al. (1999) The model contains the mapping listed here. Non-reflective: introspection and thoughtful action. Reflective: content reflection, process reflection-internal, process reflection-others, process reflection-others/internal, content reflection/process reflection-internal, content reflection/process reflection-others. Critical reflective: premise reflection.

Birney (2012) mapped indicators to four levels of reflection. Descriptive (indicator 1; see Table 1 for the description of indicators); low (indicator 2-4), medium (indicator 5-7), and high (indicator 8-12).

Poldner et al. (2014, p. 373) wrote that 'The level of a reflective essay is based on the presence of categories. For example an essay with categories description and evaluation is coded as level 2, if a reflective essay consists of description, evaluation, justification, it will be coded as level 3, etc.'.

Prilla and Renner (2014) mapped their categories to levels if at least one code of the group of codings that belong to a level is present (see Table 1 for the groupings).

The research above indicates that the descriptive categories of the models can be mapped directly to levels using certain rules.

The notion of levels might introduce a certain valuation of high levels of reflection, thus diminishing the value of lower levels of reflection. Further, because the levels are assigned to the entire text, the nuances of reflection categories that stem from the breadth dimension are summarised into only a few level categories. For example, Kember et al. (2008, p. 372) recommended for the level assignment to be based on the entire text, and not on individual parts of the text. Further, any evidence of the highest level is characterising of the entire text (Kember et al., 2008, p. 372). This implies that the highest level trumps all other levels, and thus any evidence of other levels within the text are not relevant for further analysis. Their rationale for this might be that, when considering the task of preparing aggregated statistics over a body of text in order to analyse whether a certain instruction led to improvements regarding the levels of reflective writing, a mechanism has to be established on how to assign a level to a text. Their recommendations of only considering the highest level provides a clear instruction for this case; nevertheless, it diminishes the value of the other reflection levels.

This proposed mechanism also does not record whether a lower level is missing entirely. Other conceptualisations re-frame this problem in assuming that the existence of a higher level must also include all other levels. For example, Wallman et al. (2008, p. 4) indicated that each level builds upon another. This implies that with the presence of a higher-level category, all lower level categories can be assumed to be present as well. There are several other models that share this view. For example, the model of Tsangaridou and O'Sullivan (1994, p. 20), as used in the study of McCollum (1997), starts with the description level. The next levels always assume that description is present. Such levels are 'description and justification', 'description and critique', and at the highest level, 'description, justification, and critique'.

3.1.4 *Manual reflection detection performance*

Because the goal is to not only detect reflection in writings, but also performing this reasonable well, realistic standards are important to know. This section gathers evidence on how reliable humans detect reflection.

As outlined in Section 2.2 'Models to analyse written reflection' (page 16), many of the research articles did not report any agreement/inter-rater reliability metrics. For example, Hatton and Smith (1995) used a strategy to reach agreement by discussion in order to resolve disagreements. Similarly, Williams et al. (2002, p. 8) reported that grading was performed by two instructors who negotiated agreement on a grade in case their judgement differed. Duke and Appleton (2000) stated that normal course grading procedures were followed, including cross-marking and the use of an external examiner.

However, the examined models described in Table 1 do provide this information. Table 3 provides a summary of the reported inter-rater reliability values. Table 3 shows, for each model, the number of texts analysed, chosen unit of analysis, number of coders/raters, and information provided on reliability (for a detailed description of these measures, see Section 4.2 'Evaluation criteria and metrics').

Mostly, two to three raters code the texts. The unit of analysis is either the entire text (mostly to determine the level of reflection), or smaller units of analysis, such as text segments, paragraphs, and sentences. The smaller units of analysis are mostly employed to analyse the descriptive or breadth dimension of the reflection models.

The most frequently used measure is per cent agreement, followed by Cohen's κ, Cronbach's α, and the intraclass correlation (ICC). The studies of Poldner et al. (2014) and Prilla and Renner (2014) reported Krippendorff's α values. Other statistics can be found as well. In some cases, the metric used is not clearly reported. Most models

reported inter-rater reliability values for the entire model. Some researchers provided a detailed analysis for each category, and even for each sub-category (Plack et al., 2005, 2007; Findlay et al., 2010; Poldner et al., 2014). Most papers reported only one metric. Some papers included, along with the percentage agreement, other measures (for example, Plack et al. (2005, 2007); Findlay et al. (2009, 2011); Wald et al. (2012); Mena-Marcos et al. (2013); Poldner et al. (2014); Prilla and Renner (2014)).

Author	Number of texts	Unit of analysis	Number of coders	Reliability
Sparks-Langer et al. (1990)	24 interview transcripts. 24 journals	Not clearly stated, presumably entire text	2 independent raters for the transcripts and two for the journals	Interview transcripts: 0.88% (one level difference allowed). Journals: less satisfactory
Wong et al. (1995)	45 journals. Elements: 100 reflective elements	Elements: Paragraph. Levels: student	3 independent coders	Elements: 0.5 to 0.75%. Levels 0.88%
Sumsion and Fleet (1996)	73 texts	Entire text	3 independent coders	50.00%
McCollum (1997)	202 writings on events for four journal writers	text containing meaningful events	2 independent raters	Focus dimension: 95%. Levels: not reported
Kember et al. (1999)	Coding exercise: three journals. Practical test: nine reflective papers	Coding exercise: representative sections of three journals of approximately one page length unitised by idea (two to three paragraphs). Practical test: entire paper (three to six pages)	Coding exercise: eight independent raters. Practical test: four independent raters	Coding exercise: 0.65 Cronbach's α. Practical test: 0.74 Cronbach's α
Hawkes and Romiszowski (2001); Hawkes (2001, 2006)	Face-to-face discourse: 222 chunks (comparable to a message); Computer-mediated discourse: 179 e-mail messages	Chunk (face-to-face discourse was chunked to be comparable to the messages), E-mail message	3 independent raters	Face-to-face: 0.87 Cronbach's α; Computer-medicated discourse: 0.80 Cronbach's α (reverse in the 2006 paper)
Fund et al. (2002)	Stage 1: three reflections. Stage 2: five reflections. Stage 3: 80 reflections	Sentences or part of a sentence	Stage 1: three coders. Stage 2: two external coders. Stage 3: one coder	Stage 1: Tool revision until a correlation of 98% was reached. Stage 2: Tool revision until a correlation of 98% was reached. Stage 3: Not reported
Hamann (2002)	26 code samples (out of 52 submissions)	Sentence	2 raters	Until 100% agreement was reached
Pee et al. (2002)	14 worksheets	Entire worksheet	2 raters	Per cent agreement 86%

Continued on next page

Table 3 Continued from previous page

Author	Number of texts	Unit of analysis	Number of coders	Reliability
Williams (2000) cited in Williams et al. (2002)	58 journals	Journal	3 educators	Reliability coefficient of 0.68
Boenink et al. (2004)	Responses to seven vignettes (n between ten and 15 per vignette	Response to vignette	2 independent raters	Responses to four vignettes selected from seven showed inter-rater reliability between 0.53 and 0.94 Pearson's r
O'Connell and Dyment (2004)	880 journal entries	Entry	2 external raters	Cronbach's α of 0.85
Plack et al. (2005)	43 journals out of 48	Component: Words, sentences, and paragraphs. Level: Entire journal	3 raters from initial four raters	Component: Per cent agreement (%) and Φ coefficient for pairs of raters. ICC for all raters. Levels: Per cent agreement and γ statistics, ICC for all raters. Components: Reflection-in-action: 69.8-81.4%; 0.49-0.69 Φ; 0.55 ICC. Reflection on action: 86%; 0.20-0.60 Φ; 0.41 ICC. Reflection for action: 76.7-86.0%; 0.52-0.71 Φ; 0.60 ICC. Content: 83.7-90.7%; 0.55-0.75 Φ; 0.60 ICC. Process: 81.4-93.0%; 0.10-0.69 Φ; 0.44 ICC. Premise: 81.4-90.7%; 0.62-0.81 Φ; 0.72 ICC. Return to experience: 79.1-90.7%; 0.08 Φ; 0.03 ICC. Attends to Feelings: 79.1-86.0%; 0.62 Φ; 0.28 ICC. Revaluates: 65.1-93.0%; 0.39-0.76 Φ; 0.43 ICC. Levels: 67.4-85.7%; 0.88-0.98 Φ; 0.74 ICC
Ballard (2006)	Assignments and responses to interview questions	Entire text	2 coders	Prior to the main study: First session: 90%. Second session 93%.
Mansvelder-Longayroux (2006); Mansvelder-Longayroux et al. (2007)	38 portfolios	Fragment (new learning activity)	Not reported	Cohen's κ of 0.77 for 14 out of 34 themes
Abou Baker El-Dib (2007)	20 research reports out of 100	Reflective unit: single idea or thought on particular topic or event	2 coders with scoring sheets	Cronbach's α: Problem statement 0.93, plan of action 0.81, acting 0.94, reviewing 0.83
Chirema (2007)	42 reflective journals	Elements: paragraph. Levels: student	Levels: two raters. Elements: four raters	Levels: 0.95%. Elements: 0.5 to 0.75%
Plack et al. (2007)	308 entries from 21 journals	Each journal entry (longer paragraph)	3 independent raters	Agreement and κ statistics are reported for rater pairs. ICC for all raters. Level 1: 100%; 1 κ; 1.0 ICC. Level 2: 82.8-87.6%; 0.61-0.73 κ; 0.67 ICC. Level 3: 78.2-83.4; 0.57-0.67 κ; 0.62 ICC
Kember et al. (2008)	4 journals	Entire journal	4 independent raters	Agreement: Instead of reporting, an indices table is presented and described as showing very good agreement.
Wallman et al. (2008)	56 and 126 short reflective essay (one to two pages)	Meaningful text segment and entire text	2 independent raters	Cohen's κ: 56 essays: 0.59. 126 essays: 0.65. All essays: 0.63

Continued on next page

Table 3 Continued from previous page

Author	Number of texts	Unit of analysis	Number of coders	Reliability
Chamoso and Cáceres (2009)	2432 units in reflection	Single idea that ranges from single sentence to several hundred words	Independent party who mostly agreed	93.00%
Findlay et al. (2010)	97 Journals	Not mentioned	4 independent coders	Training: Deep Analytic NRAT: 0.49 Cohen's κ (6 categories); Broad Classification NRAT 0.67 Cohen's κ (3 levels). Actual study: Deep NRAT (six categories): 0.55 Cohen's κ; Broad NRAT (three levels): 0.71 Cohen's κ. Deep Analytic NRAT (pairwise Cohen's κ): Level 1: 69.1%; 89.7 CA; 0.06-0.49 κ. Level 2: 26.8%; 75.3 CA; 0.02-0.27 κ. Level 3: 50.5%; 85.6 CA; 0.11-0.26 κ. Level 4: 89.7%; 99.0 CA; -0.02-0.16 κ. Level 5: 90.8%; 93.8 CA; -0.54-0.42 κ. Level 6: 85.6%; 93.8 CA; 0.00-0.43 κ. Broad Analytic NRAT (pairwise Cohen's κ): NR: 79.4%; 89.7 CA; 0.08-1.0 κ. Re: 75.3%; 90.7 CA; 0.08-1.0 κ. CR: 81.4%; 92.8 CA; 0.12-0.40 κ. Legend: %: Absolute agreement (all four coders agreed); CA: Consent agreement (three out four coders agreed); κ: Cohen's κ
Lai and Calandra (2010)	65 reflection writings (109-1,003 words)	Entire text	2 independent raters	88%
Bell et al. (2011)	7 journals	Line-by-line basis with meaningful text segments (phrase, sentence, number of lines, paragraph)	3 independent raters	Cronbach's α of 0.802
Clarkeburn and Kettula (2011)	Sub-sample of 263 journals	Journal	2 independent coders	84%
Findlay et al. (2011)	30 inventories	Not mentioned	2 independent coders	Deep analytic NRAT: 75.0-83%; 0.47-0.59 Cohen's κ. Broad analytic NRAT: 97.3-100%; 0.94-1.0 Cohen's κ.
Fischer et al. (2011)	110 blog posts, 45 essays	Entire writing	2 blinded coders	80%
Birney (2012)	Systematic random sampling of 27 reflective journals	Text excerpts (section of text that represents an example of one of the 12 indicators)	2 coders	Paired t-test for each of the 12 reflection indicators. Mostly no significant difference between the coders' judgements at the 0.05 level of significance
Ip et al. (2012)	6 random selected diaries	Diary	2 independent coders	95.00%
Wald et al. (2012)	Pilot 1 to 4: ten narratives each. Pilot 5: 60 narratives	Assessment of the entire narrative. Steps: First, determine criteria, then level. Second, determine overall level by considering step 1	Pilot 1 to 4: three raters each. Pilot 5: four raters	Pilot 1: 0.748 ICC. 0.899 Cronbach's α. Pilot 2: 0.455 ICC. 0.715 Cronbach's α. Pilot 3: 0.376 ICC. 0.644 Cronbach's α. Pilot 4: 0.508 ICC. 0.756 Cronbach's α. Pilot 5: 0.632 ICC. 0.774 Cronbach's α
Mena-Marcos et al. (2013)	1750 propositions (104 reflective reports)	Proposition (texts are divided into propositions. A proposition is a single meaningful predicate)	2 independent raters	Levels: 0.87 presumably Cohen's κ. Types: Not reported

Continued on next page

Table 3 Continued from previous page

Author	Number of texts	Unit of analysis	Number of coders	Reliability
Poom-Valickis and Mathews (2013)	29 cases	Text segments that contain meaning units. Identification of statements, including reflection/analysis. Reread case and assignment to the four types.	2 raters	Agreement of 67% (after discussion, 93%)
Poldner et al. (2014)	18 essays (9% of the sample)	Sentence and compound sentences	2 independent raters	Description: 85.7%, 0.71 Krippendorff's α. Evaluation: 87.3%, 0.61 Krippendorff's α. Justification: 92.4%, 0.59 Krippendorff's α. Dialogue 98.6%, 0.64 Krippendorff's α. Transfer 99.6%, 0.86 Krippendorff's α. Sub-categories: fair to good
Prilla and Renner (2014)	74 conversations and 159 comments	Single contributions to a conversation	2 independent raters	Phases/codes: Reported only for two codes: Codes 4 and 5: Krippendorff's α of 0.75 and slightly below Stage: Per cent agreement: 1. 97%. 2. 96%. 3. 80%.

Table 3: Inter-rater reliability for reflection models

Table 3 shows a wide range of reported inter-rater reliability values. In order to more easily compare the performance of the automated reflection detection with the reported inter-rater reliability values achieved by manual coders, the inter-rater reliability values are sorted into brackets or ranges. The aim is to obtain a better understanding on how frequently each of the brackets is reached by human coders. The derived ranges of inter-rater reliability values for each metrics serve as indicators on the performance of manually coding reflective texts.

Each model of Table 3 is sorted into a bracket. If a model reported several values for each of their model components, it is sorted into all those brackets that comprise the values. For example, if a model reported an agreement of 75% for one category and 83% for another, it is sorted once into the 70%-80% bracket and once into the 80%-90%bracket.

This level of aggregation seems suitable for such data. Ideally, we would like to determine commonly reached inter-rater reliability values for each category or reflective writing. However, the research results summarised in Table 3, suggest that this fine-grained aggregation is currently not possible because only a few models

reported inter-rater reliability values at the category level, and the reported metrics differ, which makes their aggregation difficult.

These brackets have been defined for the per cent agreement, Cohen's κ, Cronbach's α, and ICC. Additionally, the models reporting Krippendorff's α are discussed.

Percentage agreement values ranged from 50% to 100%. Eight papers reported a percentage agreement in the 0.5 to 0.8 bracket (Wong et al., 1995; Sumsion and Fleet, 1996; Plack et al., 2005; Chirema, 2007; Plack et al., 2007; Findlay et al., 2010, 2011; Poom-Valickis and Mathews, 2013), 11 papers reported values in the 0.8 to 0.9 bracket (Sparks-Langer et al., 1990; Wong et al., 1995; Pee et al., 2002; Plack et al., 2005, 2007; Findlay et al., 2010; Lai and Calandra, 2010; Clarkeburn and Kettula, 2011; Findlay et al., 2011; Fischer et al., 2011; Poldner et al., 2014). Eight paper reported agreement values over 0.9 (Plack et al., 2005; Ballard, 2006; Chirema, 2007; Plack et al., 2007; Chamoso and Cáceres, 2009; Findlay et al., 2010; Poldner et al., 2014; Prilla and Renner, 2014). This analysis excludes the study of Hamann (2002) because it reported the agreement value after the discussion of the raters.

Three papers reported Cohen's κ values between 0.5 and 0.6 (Plack et al., 2007; Wallman et al., 2008; Findlay et al., 2010). Two papers showed values between 0.6 and 0.7 (Plack et al., 2007; Wallman et al., 2008), and two papers noted Cohen's κ values between 0.7 and 0.8 (Mansvelder-Longayroux, 2006; Plack et al., 2007). Plack et al. (2007) achieved a Cohen's κ of 1.0 for the first level of their model.

Three papers stated Cronbach's α values between 0.7 and 0.8 (Kember et al., 1999; Abou Baker El-Dib, 2007; Wald et al., 2012), three papers values were between 0.8 and 0.9 (Hawkes, 2001; O'Connell and Dyment, 2004; Abou Baker El-Dib, 2007; Bell et al., 2011), and one paper reported a value of over 0.9 (Abou Baker El-Dib, 2007).

Three papers used ICC. Most of the reported values fall in the 0.6 to 0.7 bracket. Plack et al. (2005) reported values below 0.6 and above 0.7 for some of the categories.

The two papers that use Krippendorff's α achieved values between 0.59 (0.40 when considering the subcategories of Poldner et al. (2014)) and 0.86 (Poldner et al., 2014; Prilla and Renner, 2014).

It is notable that high percentage agreement values do not necessarily result in high Krippendorff's α or Cohen's κ values. For example, Poldner et al. (2014, p. 360) reported a percentage agreement of 99.4 and Krippendorff's α value of 0.4 for the subcategory 'Justification of Student's Choice of Situation', whereas a percentage agreement of 97.0 for the subcategory 'Description Pupils' Actions' is associated with a Krippendorff's α value of 0.77. A similar case can be found in Findlay et al. (2010) for the percentage agreement and Cohen's κ values (see Section 4.2 'Evaluation criteria and metrics' for this 'paradox').

It is also notable that those models that contain both descriptive and depth dimensions report higher inter-rater reliability values for the reflection levels than for the descriptive reflection categories (see Wong et al. (1995); Plack et al. (2005); Chirema (2007); Findlay et al. (2010, 2011)).

Several papers noted the importance of pilot testing and thorough training of the raters in order to achieve high inter-rater reliability. To quote Clarkeburn and Kettula (2011, p. 444) as an example, 'While the inter-rater reliability was satisfactory for research purposes, we highlight the intensive and collaborative process required to achieve it'.

3.1.5 *Summary*

Section 3.1 'Manual methods to detect reflection' provided an overview of the approaches used to analyse reflection. Several techniques were identified in the literature. A case was made that from these techniques, the content analysis of texts is

closest to what this thesis attempts to automate: the labelling of text units according to reflection categories.

Section 3.1.2 'Relationship between analysis units and reflection categories' investigated the question of how much textual evidence is necessary in order to assign a unit of analysis to a category. The literature revealed that frequently, a single evidence is satisfactory, but there are also other approaches that seek further substantiation until a category is deemed as present. Further, commonly, all units of analysis are coded, but other approaches exist as well.

Section 3.1.3 'Relationship between the descriptive and level reflection quality' described the mapping strategies found in those models that contained both the breadth and depth reflection qualities. In general, the literature indicated that reflection levels can be derived from the descriptive (breadth) reflection categories. This insight suggests a focus on the automated detection of reflection on descriptive reflection categories instead of level reflection models because the latter can be derived from the former by applying explicit mapping rules.

The core part of this section was Section 3.1.4 'Manual reflection detection performance'. This section informed on the performance that trained coders can achieve when classifying text units according to reflection categories. The literature revealed several types of measurements and varying inter-coder reliability values that ranged from chance agreement to perfect inter-rater reliability. The reported inter-rater reliability values were summarised for each metric. A consequence of this section is that in order to compare the performance of the automated methods to detect reflection with a human benchmark, several metrics have to be considered.

3.2 RELATED AUTOMATED METHODS

After outlining the manual method for analysing written texts with regard to reflection categories, this section describes several automated methods that can detect concepts related to reflection. Such related concepts are, for example, critical thinking, cognitive states, volition, emotions, evaluation, epistemic beliefs, and argumentation/discourse. As stated in the introduction for this chapter (see Chapter 3 'RELATED METHODS AND BENCHMARKS'), insufficient research exists on the automated analysis of writings with regard to reflection, which is the reason for this extension towards automated methods for related reflection concepts. The methods discussed have the common goal of automatically making inferences from the text to the characteristics of their concepts. This thesis shares the same goal. Although the concepts are different, the methods for related concepts can inform the selection of automated methods for reflection detection. In addition, this section informs on the performance of these automated methods. Similar to the reported performance of the manual analysis of reflection (see Section 3.1.4 'Manual reflection detection performance'), the performance of automated methods can provide an additional benchmark for the automated detection of reflection.

The discussion of automated methods is structured into three themes. First, the discussion starts with Section 3.2.1 'Dictionary-based approaches', followed by Section 3.2.2 'Rule-based approaches', and ends with machine learning approaches in Section 3.2.3 'Machine learning approaches'. After the discussion on each of the approaches, Section 3.2.4 'Automated methods performance' summarises the reported performance measures.

This order is selected to show first the two approaches based on manual vocabulary and rule set construction, before describing the research that applies a machine

learning approach that automatically derives from a set of samples, the important 'vocabulary' and 'rules' to classify text.

For each approach, work is presented that is related to the concept of reflection. The presented research is used to explain each of the three approaches, which are seen as prototypical instances of the automated methods for the classification of text.

3.2.1 *Dictionary-based approaches*

One of the earliest approaches in annotating texts was based on the use of manually compiled lists of cue words, or linguistic markers. These word lists are usually called dictionaries (concept dictionaries) or vocabulary. Each list represents a category for which the words in the list are representative.

Stone and Hunt (1963) described a tool called the 'General Inquirer', which is a program developed for automated content analysis. The system used dictionaries, each consisting of several categories (or tags) with a set of associated words. The tool annotates each word with the corresponding tag. The dictionary consists of 'first-order' and 'second-order' tags. A 'first-order' tag can only be given to one dictionary entry word. However, one or more 'second-order' tags can be assigned to a word (Stone and Hunt, 1963, p. 242). The annotated text can then be queried in order to explore the text. The queries can consist of tags and part-of-speech.

The length of the text influences word frequency, which, if used to distinguish text types based on word frequency, might lead to misjudgements. Stone and Hunt (1963, p. 249) described this problem and suggested to scale texts with a factor. Stone and Hunt (1963, p. 249) reported on a discriminative function that, in essence, describes a threshold that has to be met in order to classify text. The system also integrated a rule-based learning algorithm that can automatically find rules to distinguish texts.

A widely cited example of this approach is the Linguistic Inquiry and Word Counting Tool (LIWC) (Chung and Pennebaker, 2012), which is frequently used in research projects (Chung and Pennebaker, 2012, p. 210 ff.). The aim of LIWC is to draw an associative connection of cue words to acts of cognition. Feature words thought to be associated with psychological states are defined[2]. Pennebaker and Francis (1996) researched the link between the language used in writings and its impact on physical health and academic performance using a bank of over 60 controlled vocabularies in order to detect emotion (which is related to the category 'feelings'; see Section 2.3.2 'Common reflection categories') and cognitive mechanisms (which are related to the category 'critical stance'; see Section 2.3.2 'Common reflection categories').

LIWC serves as a prototypical example for dictionary-based approaches to detect meaning in text and is described in depth. The underlying assumption of this approach is that there exists an associative connection between cue words and acts of cognition.

LIWC's main technique is based on pattern matching the words in the text under consideration to the word lists of the software. In addition to exact pattern matching, it is possible to apply wildcards at the end of the word. For example, the word 'knowledg*' with the wildcard symbol '*' matches any word that starts with knowledg, for example, knowledg*e* and knowledg*eable* . Other forms of language processing, such as part-of-speech tagging, lemmatisation, stemming, and stop-word removal are not foreseen. Chung and Pennebaker (2012, p. 216) believed it is beneficial to work with function words, which according to them, are usually filtered by the stopword removal of other natural language processing (NLP) applications. Function words are, for example, pronouns and grammatical particles, of which LIWC makes extensive use. Indeed, a popular stop-word list excludes personal

2 See http://www.liwc.net/comparedicts.php for an overview of the dictionaries

pronouns such as 'he', 'she', 'myself', and more, and the particles like, 'because', 'since', 'thus', 'hence', etc., from full-text search[3].

LIWC vocabularies are not exclusive, and one word can belong to several categories. For example, the word 'think' belongs to the categories insight and present tense words.

The LIWC dictionaries are compiled manually, and inclusion or exclusion is based on the majority vote of three coders (Tausczik and Pennebaker, 2010). A word is included, maintained, or deleted if two out of three judges agree on the word. A second independent group of coders repeats the process.

The output of a LIWC analysis is frequency lists of the word occurrences of the dictionaries.

Compared with machine learning algorithms, Chung and Pennebaker (2012, p. 216) noted that 'NLP approaches will outperform LIWC on many classification tasks'.

It is notable that Chung and Pennebaker (2012, p. 209) stated that, 'a consistent finding is that many of the word categories that are used to reliably classify psychological states can be considered to be a part of language style as opposed to language content'. *How* people say something is sometimes more revealing than what people say.

Bruno et al. (2011) described an approach for analysing the reflective practice of learning journals using a 'mental' vocabulary. Their semi-automatic approach focusses on the detection of cognitive, emotive, and volitive words (the cognitive words are related to the category 'critical stance', emotive words to 'feelings', and volitive words can be related to an intention to do something, which is part of the 'outcome' category; see Section 2.3.2 'Common reflection categories'), thus allowing them to highlight changes in the use of these mental words over the course period.

3 http://dev.mysql.com/doc/refman/5.5/en/fulltext-stopwords.html

Instead of compiling only word lists as dictionaries, such lists can also contain phrases. Chang and Chou (2011) used a phrase detection system to study emotion, memory, cognition, and evaluation in learners' portfolios (emotions relate to the category 'feelings', cognition and evaluation to the category 'critical stance'; see Section 2.3.2 'Common reflection categories'). The system served as a pre-processor of content, thereby highlighting specific parts-of-speech (in their case, stative verbs in Mandarin because the study was conducted in Taiwan), which later helped experts to assign the automatically annotated words to four categories. Their system contained approximately 100,000 words and phrases, including word types and frequencies. They applied a semi-automatic approach by allowing the system to annotate part-of-speech, whereas researchers, teachers, and experts classified the annotated words, focusing on 'stative verbs' (i.e., intransitive verbs, causative verbs, transitive verbs, and more) into types (Chang and Chou, 2011, p. 109) that formed a 'mental' lexicon. The types of lexicon were emotion, memory, cognition, and evaluation as determined by two linguistic experts. The tool was then used to categorise writings based on word frequencies into one of three types: cognitive, evaluation, and combined..

Ullmann (2011) and Ullmann et al. (2012) used word lists of reflective keywords, question cues, and self-references, as well as premise and conclusion, learning outcome, surprise, future tense, certainty, discrepancy, and insight indicators to analyse text with regard to these categories associated with reflection.

These examples of dictionary-based approaches to analyse texts with regard to reflection-related constructs conclude this section. The above listed dictionary-based approaches were used to automatically analyse texts with regard to several aspects of cognition (Bruno et al., 2011; Chang and Chou, 2011; Ullmann, 2011; Ullmann et al., 2012; Chung and Pennebaker, 2012), emotion (Chung and Pennebaker, 2012; Chang and Chou, 2011), volition (Bruno et al., 2011), evaluation (Chang and Chou, 2011;

Ullmann et al., 2012), reflective keywords (Ullmann, 2011), premise and conclusion indicators (Ullmann, 2011; Chung and Pennebaker, 2012), questioning indicators (Ullmann, 2011), self-reference indicators, future tense indicators, certainty, discrepancy, and insight indicators (Chung and Pennebaker, 2012), learning outcome indicators (Ullmann et al., 2012), and surprise indicators (Ullmann et al., 2012).

The next section shows examples of rule-based approaches, which are frequently built on dictionaries.

3.2.2 *Rule-based approaches*

The core of a rule-based system is a set of rules that have the form of IF - THEN statements. Dictionary-based approaches can be seen as a mini rule-based system. If a text contains the cue word from dictionary X, then it is labelled as category X. In a rule-based system the rules are used together with an inference engine. This inference engine infers information based on the rules.

An example of a system based on word lists and phrases stems from research aimed at analysing discourse in texts. Discourse analysis is widely defined (Kent and McCarthy, 2012, p. 33). Usually, its focus is on written or spoken argumentation.

One stream of research on discourse analysis is applied in the area of scientific paper analysis (for example, see Teufel (1999) and Lisacek et al. (2005)).

For example, Sandor (2007) described a system that detects scientific discourse in texts (see also Sandor (2005); Sándor (2006); Sandor (2007); Sándor and Vorndran (2009); De Liddo et al. (2012)). Sandor (2007) proposed a mapping between the rhetorical function (e.g., research background, author contribution) and meta-discourse.

'Metadiscourse is a cover term for self-reflective expressions used to negotiate interactional meanings in a text, assisting the writer [or speaker] to express a viewpoint and engage with readers as members' (Hyland, 2005, p. 37).

The concept-matching model for meta-discourse of Sandor (2007) consisted of constituent concepts and markers. Markers are realisations of a constituent concept. Several constituent concepts form a rhetorical function. The markers that reflect a constituent concept have to be in a syntactically linked co-occurrence relationship.

The approach for constructing such a system consists of three steps (Sandor, 2007, p. 105) conducted by experts. First, large corpora are analysed to derive a small number of words seen as related to constituent concepts. Second, the list of words is extended in order to generate a representative list of words for the constituent concept. Third, based on the constituent, concepts rules are generated and tested. The rules combine those constituent concepts that realise a rhetorical function.

With regard to the performance of the approach, Sandor (2005) stated that it is difficult to evaluate recall, whereas precision can be high (almost 100%). They noted that the dictionaries have to be constructed manually because there is no automatic way of constructing these.

Other examples of a rule-based system are derived from the research area of dialogue analysis (for example, see Graesser et al. (2012); Dessus et al. (2009); Katz et al. (2000)). Erkens and Janssen (2008) devised a system that aims to code dialogue acts to 'determine the communicative function of messages in online discussions by recognizing discourse markers and cue phrases in the utterances' (Erkens and Janssen, 2008, p. 487).

The system was first used for manual coding in the 'Verbal Observation System' (Erkens and Janssen, 2008, p. 450), and later implemented into an automatic system for the detection of dialogue acts, called 'Multiple Episode Protocol Analysis' (Erkens and Janssen, 2008, p. 453). The system distinguishes five elements of dialogue acts, which in turn comprise 29 sub-elements in total. The elements are argumentative, responsive, elicitative, informative, and imperative dialogue acts (argumentation

relates to the category 'critical stance'; see Section 2.3.2 'Common reflection categories').

Here, the underlying assumption is that discourse markers or clue phrases are generally used by people to indicate to others the intended purpose of the utterance. The amount of discourse markers is seen as limited, and therefore, in principle, codifyable by computers.

The system applies a two-step approach. First, the utterances are segmented. Such segmentation considers punctuation, connective words, discourse markers, and certain exceptions (Erkens and Janssen, 2008, p. 453). Afterward, the segments are annotated according to their dialogic acts. The system uses pattern matching with if-then rules. An example is[4]: If 'so' is in the protocol field, then code variable V6 as the 'Conclusion'. This detects discourse marking words, phrases, idioms, or partial phrases. In case a message cannot be coded, it is marked with a special annotation. Approximately 10% of the messages are assigned to this annotation (Erkens and Janssen, 2008, p. 454), which can then be manually annotated. Erkens and Janssen described the use of exception rules in order to prevent one dialogue act indicator to be assigned to more than one element (Erkens and Janssen, 2008, p. 454).

Erkens and Janssen (2008, p. 456; 460ff) compared manually coded texts with the automatically coded texts of two random segments of 500 chat messages. They used one human coder experienced in the dialogue act coding system. A total of 29 categories were coded. There was disagreement on 210 messages (21%), from which 106 messages belonged to the comprehensive 'remaining cases' category. Agreement over all elements was 79% (85.6% if omitting the comprehensive category) and Cohen's κ was 0.75 (0.84) (Erkens and Janssen, 2008, p. 461). The agreements varied for each element from 0.04 to 0.84 on the lower end of each category, and 0.71 to 1.00

4 http://edugate.fss.uu.nl/mepa/prodrules.htm

on the higher end. The individual values for each dialogue act annotation were not presented.

Ullmann et al. (2012) proposed a rule-based system to annotate texts whether they are reflective or not. The rule-based system chained together dictionary-derived facts with a set of rules to infer the reflectiveness of texts. The system combined a dictionary-based approach with a rule-based approach.

Both dictionary-based and rule-based approaches rely on the manual construction of dictionaries or rules. Once these are modelled, the software can automatically analyse texts.

The examples outlined above should provide sufficient introduction to rule-based approaches. Both approaches have been explored in other works of the author of this thesis (Ullmann, 2011; Ullmann et al., 2012). The next section discusses approaches based on machine learning. This approach generates these models automatically from data.

3.2.3 *Machine learning approaches*

The approaches outlined above rely more or less on pattern extraction methods, mapping predefined words collected in dictionaries to categories, and rule-based systems, which often extend dictionary-based approaches. The next type relies on machine learning algorithms (Sebastiani, 2002), especially text mining algorithms (Gupta and Lehal, 2009), to classify content. The approaches are not mutually exclusive. Machine learning approaches also utilise dictionaries, for example, to reduce feature space. The main advantage of machine learning is seen in its ability to automatically detect classification rules, compared with human-generated rules, as demonstrated in the above outlined approaches of Ullmann et al. (2012), Erkens and Janssen (2008), and Sandor (2007, p. 100), for example.

Research that applies machine learning algorithms in order to gain insight on learning grows continuously. Several reviews that discuss machine learning in the area of educational science have been published (Romero and Ventura, 2006, 2007, 2013; Baker and Yacef, 2009; Baker, 2010; Ferguson and Shum, 2012; Clow, 2013; Bienkowski et al., 2012; Baker and Inventado, 2014; Papamitsiou and Economides, 2014).

On a conceptual level, machine learning methods can be distinguished into supervised or unsupervised methods, methods that work on quantitative and qualitative data, or both. In this section, the focus is on supervised methods successfully applied to qualitative, unstructured data, such as text. The main difference between supervised and unsupervised machine learning is that in the former case, the data are labelled and the goal is to predict labels for new, unseen text units. Unsupervised methods work without labels. Their main goal is to detect structure in data for exploratory purposes (see also the discussion in Section 4.5 'Machine learning algorithms').

Regarding the application of machine learning for education on qualitative, unstructured textual data, a large body of research exists in the area of e-assessment (e.g., Kalz et al. (2014)), and especially, in the area of automatic essay assessment (Page and Paulus, 1968; Page, 1968; Hearst, 2000; Landauer, 2003; Shermis and Burstein, 2003; Wild et al., 2005; Attali and Burstein, 2006; Dikli, 2006; Alden Rivers et al., 2014; Jordan, 2014; Shermis, 2014). Other strands of research are described as applied natural language processing, which has many connections to the field of education (for an overview, see McCarthy and Boonthum-Denecke (2012)). Furthermore, machine learning is applied in the research area of discourse analysis (for an overview, see Dessus et al. (2009); Ferguson and Shum (2011); Dascalu (2014)).

Similarly to the previous section, the discussion in this section focusses mainly on the techniques applied to problems related to the automated detection of reflection. The

section is divided into two subsections, the first reports on research based on the use of single machine learning algorithms (see Section 3.2.3.1 'Single classifiers'), and the second describes research that used multiple machine learning algorithms to classify text (see Section 3.2.3.2 'Multiple classifier').

3.2.3.1 *Single classifiers*

This section presents research that applied a single machine learning algorithm to their classification problem. The machine learning algorithms are artificial Neural Networks, Naïve Bayes classifier, and Support Vector Machines (SVMs).

McKlin (2004) employed artificial Neural Networks to build classifiers for the automated classification of discussion posts according to the four categories of cognitive presence, namely 'triggering events', 'exploration', 'integration', and 'resolution'. According to Garrison et al. (2001, p. 11), cognitive presence comprised '(...) higher-order knowledge acquisition and application and is most associated with the literature and research related to critical thinking'. Critical thinking relates to the category 'critical stance' of the common reflection categories described in Section 2.3.2 'Common reflection categories'.

McKlin (2004, p. 45) outlined that the application of Neural Networks requires a large number of training data in order to achieve reasonable results. Six coders coded 1,200 messages to build the training corpus for the artificial neural network algorithm. The coders received training for this coding. Training continued until a pairwise Cohen's κ of 0.7 was reached. Most raters had 90 messages for training (McKlin, 2004, p. 129 ff.). After training, each coder rated 300 messages that were selected by course (200 messages) and from all courses (100 messages). The latter set of 100 messages had to be rated by all coders, and served to calculate inter-coder agreement. The mean Cohen's κ was 0.6 (percentage agreement, 74%), ranging from 0.49 to 0.74 (McKlin, 2004, p. 86 ff.).

McKlin (2004) described that the Neural Network algorithm expected numeric input. The numeric input was based on the frequencies derived with a dictionary-based approach. These dictionaries were compiled from the 182 categories of the 'The General Inquirer' (Stone and Hunt, 1963) and 37 manual categories (McKlin, 2004, p. 58).

McKlin (2004, p. 58) described an example on the sentence '*YES*, I had to look *UP* to see the icon'. The words 'yes' and 'up' belong to the dictionary 'positive' words. This message is then coded as '2' on the dictionary category 'positive'.

A total of 1,100 messages were used as the training set, and 100 messages were used for testing. For the first experiment, the best 102 predictors were used as input parameters. The Neural Network used two hidden layers and five outputs for the five elements of the model used in the first experiment (McKlin, 2004, p. 194). The second experiment used a three-layer network with 40 inputs and four outputs (McKlin, 2004, p. 195).

McKlin (2004) described two evaluation cycles. The results of the first cycle with the best machine learning model showed a percentage agreement of 71% and a κ of 0.52 with human judgements (McKlin, 2004, p. 89).

Because of the low agreement, a second cycle was started that refined the training data set. The reasoning behind this is that it was assumed that the accuracy of the Neural Network model would increase if the agreement between human coders were higher (McKlin, 2004, p. 90). This cycle increased the coder agreement. The mean pairwise Cohen's κ for the 100 messages coded by all four raters was 0.85 (mean percentage agreement 90%), ranging from 0.79 to 0.91 (McKlin, 2004, p. 95).

Based on the refined training data, Neural Network models were trained. From the 1,600 coded messages, 125 were removed because they were difficult to code. A total of 1,180 messages were used for training the Neural Network, whereas 295 messages formed the test set. Again, several Neural Network models were created. The models

were used to code the 100 messages for comparison with the human coders. The model with the best overall reliability was maintained (McKlin, 2004, p. 97). The highest Cohen's κ was 0.70 (81% percentage agreement).

The best model had 40 inputs. The three most discriminative features were word count, questions, and the name of the person from the class (McKlin, 2004, p. 97 f., p. 196).

Corich et al. (2006) described a system (the Automated Content Analysis Tool, or ACAT) that classifies texts according to the cognitive presence model (Garrison et al., 2001) and the model of Perkins and Murphy (2006). A Naïve Bayes classifier was used to distinguish content according to categories of critical thinking.

Corich (2011) reported two research cycles for the first model. The data used in the first research cycle consisted of 74 posts, which in turn resulted in 484 sentences (the unit of analysis were sentences). A total of 50 sentences served as training for the two coders. For the cognitive presence model with four elements, a Holsti's CR of 0.82 and a Cohen's κ of 0.76 was reported (Corich, 2011, p. 126).

These data, labelled by the coders, were then used to train the model. Best results in comparison with human judgements were a Cohen's κ of 0.68 and Holsti's CR of 0.71 (Corich, 2011, p. 151).

For a second study, 80 posts with a total of 310 sentences were analysed. Corich (2011, p. 162) described that from a total of 462 sentences 152 sentences that were of a social nature or that did not contribute to the discussion topic were removed. The agreement between coders was 80% (Holsti's CR) and 0.76 (Cohen's κ).

The automated analysis of the new data was based on the training corpus of the previous study. Corich (2011, p. 165) reported that the agreement between the automated system and the coders were lower, with the best results being 0.43 (Holsti's CR) and 0.41 (Cohen's κ). A re-training of the data set resulted in 0.67 Holsti's CR and 0.65 Cohen's κ (Corich, 2011, p. 167).

The theoretical model of Perkins and Murphy (2006) was used for another experiment. This model is especially designed to measure critical thinking at an individual level (Corich, 2011, p. 180f.). The 800 units of the previous forum posts were re-coded. A total of 400 units were coded with two coders. Disagreements were removed in mutual understanding. This data served as the training data for the next experiment.

The 142 students posts contained 436 sentences. A total of 148 sentences were removed because they were either of a social nature or off-topic. Therefore, 288 sentences formed the test data. The reliability between the automatically coded results and the human coders reached 0.67 (Holsti's CR) and a Cohen's κ of 0.65 (Corich, 2011, p. 187).

Kovanovic et al. (2014) used an SVM algorithm to detect the four elements of cognitive presence of the model of Garrison et al. (2001). Their data set consisted of 1,747 messages. Two coders rated each message according to the four categories of cognitive presence. They achieved very high inter-rater reliability (per cent agreement of 98.1% and a Cohen's κ of 0.974), which is much higher that the studies reported above. Their data set consisted of 308 messages classified as 'triggering events', 684 'exploration', 508 'integration', 107 'resolution', and 140 messages labelled as 'other' (messages in the 'other' category were not included in the machine learning training phase).

As resampling strategy, Kovanovic et al. (2014) chose ten-fold cross-validation. The tuning parameters of the linear SVM algorithms were set at a constant value, which may have limited the performance of the classifier. It is notable that they tested the influence of feature engineering on the performance of the classifier. In total, 14 different feature sets were tested. These included, for example, unigrams, n-gram variants, part-of-speech n-grams, and dependency triples. They excluded those features that occurred in the entire data set fewer than ten times. The model based on

unigrams reached a Cohen's κ of 0.364, whereas the model based on back-off trigrams achieved a Cohen's κ of 0.410 (the data with this feature set had the most additional features).

3.2.3.2 *Multiple classifier*

Dönmez et al. (2005) used the Minorthird text-learning toolkit (Cohen, 2004) to classify text segments according to seven dimensions of argumentative knowledge construction (Weinberger and Fischer, 2006). Such dimensions are: epistemic activity, micro-level argumentation, macro-level argumentation, social modes of interaction, reaction, treatment check dimension, and quoted dimension (quoted means whether the text was quoted as a reply).

1,255 pre-coded text were used as training data for the classification algorithm. A first test applied K-Nearest Neighbour as classifier with ten-fold cross-validation. The κ values ranged from 0.35 to 0.81. The categories closest to this research were the dimension epistemic with a κ of 0.51 and micro-level and macro-level argumentation, which both had a κ of 0.54. The second experiment was based on voted perceptron learning algorithm. The reported κ values ranged between 0.49 and 0.98, with 0.49 for the dimension epistemic, 0.76 for the micro-level argumentation, and 0.67 for the macro-level argumentation.

Rosé et al. (2008, p. 255) reported results that applied several classification algorithms (Naïve Bayes, SVMs, and Decision Trees) to the same coding schema of Weinberger and Fischer (2006) reported in Dönmez et al. (2005). Cohen's κ ranged from 0.49 to 0.91 (epistemic dimension was 0.49, micro-level argumentation was 0.6, and macro-level argumentation was 0.7). They used two programs, TagHelper (Rosé et al., 2008; Dönmez et al., 2005) and SIDE (Mayfield and Penstein-Rosé, 2010; Kang et al., 2008), which are mainly built around the machine learning algorithms of the WEKA toolkit (Hall et al., 2009). The tools provide the means to reduce the feature space, thus offering

features such as length, punctuation, unigrams, bigrams, part-of-speech, pre-defined dictionaries, and an interface to define queries.

Dönmez et al. (2005) and Rosé et al. (2008) did not describe the inter-rater reliability of the dataset. Weinberger and Fischer (2006) reported the inter-rater reliability of the framework to analyse argumentative knowledge construction. The Cohen's κ for the epistemic dimension was 0.9, for the argument dimension it was 0.78, and for the social modes of co-construction dimension it was 0.81.

Whereas the previously mentioned approaches did not research reflective thinking per se, the following work aimed at identifying reflection with machine learning algorithms. Modupeoluwa (2011) compiled a corpus of blog posts on the reflection of work received from a medical school, web blog posts (posts about students' interest in university courses), and reports of past students' computing projects (especially considering the section on student project reflection).

Four classifiers were applied using the WEKA toolkit (Hall et al., 2009). It was concluded that the Naïve Bayes Classifier and SVM classifier yield the best results (compared with the Zero-R classifier and the J-48 Pruned Classifier). Cohen's κ was not reported.

The texts were classified as reflective based on the label of the text section. The project report did not contain information as to whether the texts were actually reflective. A manual content analysis of the chosen texts was missing. This would have helped to assess the validity of this work because it would have ensured that the texts were actually reflective or non-reflective.

3.2.4 *Automated methods performance*

Several of the machine learning approaches focussed on the model of Garrison et al. (2001). Corich (2011) reached a Cohen's κ of 0.68 with a Naïve Bayes classifier. McKlin

(2004) reached a Cohen's κ of 0.70 using Neural Networks. The work of Kovanovic et al. (2014) reported a Cohen's κ of 0.41.

The work of Dönmez et al. (2005) and Rosé et al. (2008) applied several machine learning algorithms in order to automatically analyse collaborative learning processes. They reached Cohen's κ values ranging from 0.53 to 0.76 on related concepts of reflection. It is notable that these two papers showed that, by experimenting with different classification algorithms, improvements for certain categories could be achieved. For example, the cascaded binary classification algorithm had a κ of 0.76 for the category micro-level of argumentation (Dönmez et al., 2005, p. 132), whereas the SVM algorithm only achieved a κ of 0.60 (Rosé et al., 2008, p. 255). SVM achieved slightly higher κ values on the macro-level argumentation. These results tend to support the approach taken in this thesis, which is to test several classification algorithms in order to evaluate their differences with regard to performance.

3.2.5 *Summary*

The high Cohen's κ scores seem encouraging for further research in this area, especially given that some of the categories used in such research were related to the reflection categories (see Section 2.3.2 'Common reflection categories'). Argumentation can be seen as part of the category 'critical stance' because being critical about something might be expressed in the form of a line of argumentation. The epistemic dimension contained categories such as the 'construction of the problem space' or 'conceptual space', and the linking of both. This is related to the category 'description of an experience', which is a recollection of what occurred. Describing a past experience can be the problem space on which a person starts to reflect. The work of McKlin (2004) and Corich (2011) was based on the critical

thinking model of Garrison et al. (2001). The category 'critical stance' of the model of reflection is related to critical thinking.

Machine learning approaches are based on pre-annotated data sets used to train the classifier. From the outlined research, it seems beneficial for the generation of robust predictive models for the training set to be of a representative size with many instances for each category.

The intended use case for these predictive models was to help researchers with the annotations of texts and text segments for content analysis. For example, ACAT contains a quantitative content analysis (QCA) training module with which users can train a model for their content analysis. This model has to be created in advance with a model management tool (Corich et al., 2006).

The research on machine learning algorithm also outlined the importance of dictionary-based approaches. For example, McKlin (2004) used a dictionary-based approach to control the feature space for the Neural Network algorithm. Machine learning algorithms, dictionary-based approaches, and rule-based approaches can support each other.

The benefit of dictionary-based and rule-based approaches is that the generated word lists and rules are manually constructed by experts, and thus, they are well understood. In essence, it is a codification of the patterns derived from the experience of domain experts. As such, they can be used to explain why the system arrived at a certain conclusion. In addition, these systems can achieve high predictive power. For example, the rule-based system of Erkens and Janssen (2008) reached a Cohen's κ of 0.75.

4

METHODOLOGY AND RESEARCH DESIGN

This chapter discusses the methodological considerations that influenced the choice of the selected methods and design decisions of the research design. This chapter is divided into six major parts.

Section 4.1 'General methodological considerations' provides the rationale for choosing a quantitative research approach using machine learning as its main component. Consequently, Section 4.2 'Evaluation criteria and metrics' outlines the evaluation criteria of the study and provides an overview and discussion of relevant measurements in the context of this research. Thereafter, two other general design decisions are motivated. First, Section 4.3 'Unit of analysis' provides guidelines on the choice of the analysis unit and justifies the chosen unit. The choice of the analysis unit is important because it determines the basic unit for each observation. This is followed by a discussion on sampling strategies. The choice of sampling strategy influences the generalisability of the research results. Section 4.4 'Sampling' introduces several sampling plans and discusses their benefits and limitations.

Subsequently, Section 4.5 'Machine learning algorithms' follows the discussion on which of the many machine learning algorithms are good candidates for the detection of reflection in writings. Along with the discussion on the choice of machine learning methods, this section outlines their concrete implementation.

These sections provide all the information necessary for the description of research design. The research design outlined in Section 4.6 'Overview of research design' contains two major parts. First, Section 4.6.1 'Dataset generation process' describes the

process used to generate the dataset on which the machine learning algorithms are trained and tested. Second, Section 4.6.2 'Research design' outlines the research design and its methods used to answer the research questions.

4.1 GENERAL METHODOLOGICAL CONSIDERATIONS

Before outlining the research design, the following paragraphs summarise the general considerations with regard to the methodology. The discourse starts with the question whether a qualitative or quantitative study suits the aim of this thesis.

Many research studies that analysed written reflection used an interpretative qualitative approach. Examples for this type of research are listed in the cited literature in Section 2.2 'Models to analyse written reflection' (on page 16). Many of the models for reflective writings were derived by an interpretative, close reading of the original literature on reflection. This type of research is useful for applying more precisely the theory of reflection to the context of the analysis of written reflection. In this thesis, the aim is different: it is to compare the performance of automated methods to detect reflection with a human baseline. A quantitative approach is most suitable for this research because it quantifies the difference between generally expected human performance and the performance of automated methods. This is also the reason that the literature review on the models of reflection is tailored to research studies that apply quantitative content analysis. This type of research produced statistics that serve as the measurement for comparison.

In addition to quantitative content analysis, reflection is assessed with various other quantitative methods (see Section 3.1 'Manual methods to detect reflection' on page 57). Most prominent are questionnaires, but also tests that assess reflective judgement. Once the data are captured, the results can be generated automatically. However, these

quantitative methods are used to gather structured data in order to evaluate reflective thinking skills. They do not allow directly inferring reflection qualities based on text.

Although this research is based on textual data often associated with qualitative methods of inquiry, the type of study used to answer the research question is quantitative. The predicted values of the machine learning model are compared with the actual, true value of the dataset. Both the predicted and actual values are categorical data that serve to infer several measurements that indicate the performance of the machine learning algorithms on the problem of reflection detection. This qualifies this research as a quantitative study.

After de-emphasising qualitative studies and other quantitative methods as candidate methods for this thesis, the question is why this thesis assumes that automated methods can be used to detect reflection. Automated methods only work if there are regularities, patterns, or structure that can be codified. Thus, with the thesis that automated methods can be used to detect reflection comes the claim that there are regularities of expressing reflection. The proposal of using automated methods to detect reflection is built on the hypothesis that there are regularities of how reflection is expressed in texts.

The assumption that such regularities exist is not unwarranted. Supporting evidence can be found in the literature of the analysis of reflective writings. The research outlined in Section 3.1.4 'Manual reflection detection performance' indicates that trained humans can reliably classify text with regard to model categories; however, a stronger argument for this assumption is the evidence of regularities at the textual level, and there are indicators for this. Early on Hatton and Smith (1995, p. 42), based on their experience with coding reflective texts, the authors remarked that they found language patterns that helped them code dialogic reflection. Further, Fund et al. (2002, p 491) described patterns related to the reflection categories and the coordination of idea units. Poom-Valickis and Mathews (2013, p. 423) recorded

keywords for each reflection category that helped them find categories with similar connotations. Hawkes and Romiszowski (2001), Hawkes (2001), and Hawkes (2006) indicated a link between discourse markers and the reflection model of Sparks-Langer et al. (1990). Further, evidence of linguistic patterns originated from the research on reflection analysing text with systemic functional linguistics (Shaheed and Dong, 2006; Luk, 2008; Reidsema and Mort, 2009; Forbes, 2011; Ryan, 2011; Birney, 2012; Ryan, 2012; Wharton, 2012; Ryan, 2014), and from our research using a dictionary-based approach (Ullmann, 2011) and a rule-based approach (Ullmann et al., 2012).

The next question is why machine learning methods? The research on automated methods that investigated the detection of related thinking skills (see Section 3.2 'Related automated methods') shows two other types of methods, in addition to machine learning approaches, called dictionary-based and rule-based approaches. These are good candidate techniques, especially when combined, for the automated detection of reflection and were explored in Ullmann (2011) and Ullmann et al. (2012).

Dictionary-based approaches have advantages in quickly producing frequency tables for each category, classified with the help of dedicated dictionaries. The dictionaries can be inspected, which aids the understanding of the assignment of categories to texts.

Rule-based approaches have been applied successfully for classifying text chunks, such as sentences or paragraphs. They are especially successful in areas where the variability of expressing the searched concept is limited or extremely well understood.

Both techniques allow manually developing proxies for reflection in texts. However, the development of these proxies is mostly manual work. This thesis improves the automation of detection by evaluating algorithms that can build models for these proxies on their own. Machine learning algorithms have shown great potential for the classification of text. The research outlined in Section 3.2.3 showed promising results on related concepts of reflection.

After outlining the reasons for choosing a quantitative study type, and for choosing machine learning algorithms to detect reflection, the next section outlines several other general considerations before describing the research design, its processes, and methods. The next section describes the three evaluation criteria of reliability, validity, and objectivity. It follows a summary of measurements derived from the literature review of related methods, an overview of benchmarks, and the selection of measurements suitable for this research.

4.2 EVALUATION CRITERIA AND METRICS

This section outlines several criteria that are used to evaluate the quality of the research method. They are the quality criteria of reliability, validity, and objectivity.

Further, this section discusses the measurements used to evaluate the performance of the methods related to this research (see Chapter 3 'RELATED METHODS AND BENCHMARKS').

One of the three classic quality criteria is reliability, which expresses the extend to which measurements produced by one measurement instrument can be reproduced by other measurement instruments (Gwet, 2012, p. 9). Reliability ensures that the results can be replicated under similar conditions.

Section 3.1.4 'Manual reflection detection performance' shows research that reported the measurements of inter-rater reliability. This measure is important because high inter-rater reliability expresses a certain confidence that the method for classifying the proposed model components for reflective writing can be replicated by other researchers.

This confidence is often encapsulated in a common approach for measuring reliability on a random subset of the original data. If the reliability is high, the remaining data can be evaluated by a single coder only. High reliability between

coders is seen as an indicator that the ratings of one coder are the same as the ratings of another coder. They both provide the same information. This implies that not all data have to be evaluated by several coders (Stemler and Tsai, 2008, p. 37f.).

The measures outlined in Section 3.1.4 'Manual reflection detection performance' and Section 3.2.4 'Automated methods performance', and summarised in this section, are important when determining the quality of the machine learning models. There, the classification results of the machine learning model are compared with cases known to be correct. In a sense, inter-rater reliability is calculated between two coders: the first coder provides the gold standard for true labels, and the second coder is the classification model that predicts these labels for each analysis unit of the test dataset.

Validity is a measure to ensure that what is measured is actually that which was intended to be measured (Gwet, 2012, p. 10). For example, does the automated method to detect reflection actually measure reflection? Content analysis, compared with psychological testing (Association et al., 1954, 1993), has its unique challenges to provide evidence on the validity of the sought construct. The measurement of validity is an iterative process that attempts to find evidence from imperfect and indirect methods to build stronger arguments on validity (Krippendorff, 2012, p. 333).

Reliability is a necessary, but not sufficient, criterion for validity. The connection is that 'unreliability limits the chance of validity' (Krippendorff, 2012, p. 268), and 'reliability does not guarantee validity' (Krippendorff, 2012, p. 269). Krippendorff (2012, p. 270) warned that a common experience among content analysts is that 'in the pursuit of high reliability, validity tends to get lost'. Poldner et al. (2012, p. 31 f.) highlighted in their review of quantitative content analysis of reflective writings that none of the studied articles contained information on the empirical validity of their category schema.

Objectivity is usually interpreted as how independent or unbiased the results are from external conditions. It is an important criterion to achieve replicable results by

other researchers. Researchers can ask the question (Kassarjian, 1977, p. 9): 'Can other analysts, following identical procedures with the same set of data, arrive at similar conclusions?' In order to achieve this, several considerations can be taken into account.

Researchers can ensure the results to be independent from the person(s) conducting the research. For example, researchers should not take part in the course to be evaluated because they might provide instructions that influence the writing style of the participants towards the researchers' expectations (avoiding experimenter's bias).

Using independent raters, i.e., those who did not take part in the research preparation and experiment, fosters replicability. A coding book can be used to explicate all rating decisions. In addition, if the coders do not know the participants, only the text might be evaluated, thus making it less likely for their coding decisions to be biased by the knowledge of the text's writer.

Another point to consider is that the coders should rate texts independently, against the practice of discussing difficult coding decisions in order to achieve mutual agreement.

The criterion of objectivity is important for the data generation step. Attention should be paid for the raters to code the data independently, based on explicit coding instructions. In addition, coder training has to be described in detail in order for other researchers to replicate the results. The research described in Section 3.1.4 'Manual reflection detection performance' only provided detailed information on this step for some cases. This resonates with the finding of Poldner et al. (2012, p. 32), which was that none of the 18 reviewed papers included information on the necessary amount of training.

There is no risk of compromising objectivity when using machine learning models to predict the label for new text units because these models deterministically predict the same outcome for the same input.

Reliability and validity can be determined empirically. Several measures of reliability and validity exist. The following sections highlight the important metrics used in the research work of the following related methods sections, as well as the metrics used later to assess the quality of the automated detection of reflection.

In particular, in the context of the research on the manual analysis of reflection agreement, reliability, and correlation metrics are used to evaluate the quality of the manual content coding. As outlined in Section 3.1.4 'Manual reflection detection performance', the most frequent measures are the per cent agreement, Cohen's κ, Cronbach's α, Intraclass Correlation Coefficient (ICC), Krippendorff's α, and Pearson's r (Pearson product-moment correlation coefficient).

The reported measures to compare human and machine coding outlined in Section 3.2.3 'Machine learning approaches are per cent agreement, Cohen's κ, and Holsti's coefficient of reliability (CR).

The following paragraphs discuss all these metrics. In addition, this section explains those metrics that are important for describing the performance of the machine learning algorithms.

Per cent agreement is one of the simplest and most common measures of agreement.

$$\text{Per cent agreement} = \frac{\text{Number of agreements}}{\text{Number of all cases}} * 100 \tag{1}$$

The average pairwise per cent agreement can be used for the case of three or more coders. A variation is the calculation of agreement with tolerance. For example, tolerance by one means that two raters still agree if their coding differs by one.

There is a debate on reporting only the per cent agreement when describing the reliability of a study. The argument is that it does not consider chance agreement and can lead to an overestimation of the agreement, especially when there is an imbalance

in the category frequency (Lombard et al., 2002, p. 590) (see an example of this in the discussion on per cent agreement in the context of other reliability indices in Section 3.1.4 'Manual reflection detection performance' on page 70). However, there are cases where the raters are extremely consistent with their rating (high agreement), but the probability of the reliability indices is low (e.g., low Cohen's κ). This is also often the case where the data are imbalanced. A reliability index proposed for such situations is Gwet's AC_1, which is discussed in the context of other reliability measures below.

The benefit of the per cent agreement is that it can be interpreted intuitively and it is easy to calculate. When using per cent agreement, it is recommended to also report a metric that accounts for chance agreement (Lombard et al., 2002, p. 600).

Holsti's CR is a modified version of the per cent agreement, and as such, it does not consider chance agreement (Rourke et al., 2001). Holsti's CR was described ten years earlier, and it is referred as Osgood's coefficient (Krippendorff, 2004b, p. 417).

Cronbach's α (Cronbach, 1951; Cronbach and Shavelson, 2004) is a measurement of internal consistency frequently used in the item analysis of questionnaires to create summary scales. Pearson's r ranges from -1 to 1. A correlation of 1 indicates a strong positive correlation, whereas a correlation of -1 indicates a strong negative correlation. A correlation of 0 indicates no correlation. This is related to Pearson's Φ coefficient (aka r_ϕ coefficient), which calculates the strength of the association between two binary variables. The correlation coefficients are often used to empirically assess the validity of a measurement tool. Cronbach's α was reported as a reliability measure in the research about the manual content analysis of reflective writing (see Section 3.1.4 'Manual reflection detection performance'). However, Krippendorff (2012, p. 307), Krippendorff (2004b, p. 428f.), and Müller and Büttner (2006) noted that Cronbach's α and Pearson's r are measures of association and not of reliability.

After the description of the per cent agreement and several association measures, the focus now shifts to indices of inter-rater reliability (for an introduction to reliability measures, see Krippendorff (2012, p. 267-328) and Gwet (2012)).

Frequently cited in the related methods section of this thesis (see Section 3.1.4 'Manual reflection detection performance' and Section 3.2.4 'Automated methods performance') is Cohen's κ, a chance-corrected agreement measure (Cohen, 1960) that aims to overcome the limitations of the per cent agreement by removing the proportion of agreements that can be expected by chance. Cohen's κ can only be used for two raters, but several extensions have been proposed for multiple raters (see Fleiss et al. (2003, p. 610 ff.)).

There is not one ICC, but many different forms of them. Shrout and Fleiss (1979) outlined six models. Several considerations have to be made when choosing the right model for the given study. The guidelines are provided by Shrout and Fleiss (1979). Müller and Büttner (2006) provided a decision tree to select the appropriate ICC model. ICC(1,1) is proposed for the situation where each unit is rated by multiple raters. ICC(2,1) can be used when all units are rated by the same set of raters. Both raters and units are selected randomly. The conditions for ICC(3,1) are similar to the conditions for ICC(2,1), with the exception that the raters are not selected randomly (Gwet, 2012, p. 151 ff.). The number '1' that follows the model number signifies that this is a model that considers the ratings of the individual raters. In addition, all three models exist for case where the averages of the raters are used instead of the individual ratings, and they are denoted as ICC(1,k), to ICC(3,k). The 'k' indicates that it is the average and not the individual model of k raters. They are useful for situations where the ratings of the individual raters can be seen as unreliable (Gwet, 2012, p. 151 ff.).

Krippendorff's α is a statistic for reliability, and it usually spans between zero and one. A value of zero represents coding that would be expected by chance, whereas a

value of one represents perfect reliability. A value smaller than zero indicates a structural error in the data (for example, if coders systematically code differently, it might be because of a misunderstanding in the coding process).

Krippendorff's α is described as a statistic that considers chance agreement, it is applicable to different sample sizes, it scales, it is robust in the case of missing data, and multiple coders are allowed (see Krippendorff (2004a, p. 787) and Krippendorff (2012, p. 278)).

Another interesting property is that the rater is generalised out of the equation. Krippendorff (2012, p. 282) outlined that the individuality of the raters is not necessary for a reliability index to be considered, because only the aggregated evaluation for each unit counts. This is important for the context of the process that generates the dataset. There, the ratings are originated by different individuals, unlike the traditional case where the same person rates all items once. This property of Krippendorff's α does not require coders to complete all questions. They can stop at any time.

Another metric with similar beneficial characteristics as Krippendorff's α is the AC_1 coefficient proposed by Gwet (2008, 2012). AC stands for Agreement Coefficient. The subscript 1 is used to differentiate the AC_1 from the AC_2. The AC_1 considers first-level agreement (total agreement) and the AC_2 considers second-level agreement (partial agreement) (Gwet, 2012, p. 78 f.). In the case of nominal data, κ and α statistics sometimes produce extremely low reliability values, although high agreement is present. Gwet (2008) positions his AC_1 coefficient in this context to resolve this 'paradox' of low reliability values, although the data indicate high agreement (Gwet, 2012, p. 36ff).

In addition to deciding the statistics to measure reliability, it is also important to discuss the level of reliability that is acceptable. Although there cannot be a definite

answer because the level relies on the particularities of the specific research, some benchmarks for reliability measures have been proposed.

Krippendorff (2012, p. 325) recommends the following levels for Krippendorff's α based on research in social science:

- α values between 0.667 and 0.800 for tentative conclusions
- α values over 0.800 are reliable

However, these values have to be seen from the perspective of the possible research consequences. For example, more strict guidelines should be enforced for research with high cost for wrong conclusions, and less strict for the case of exploratory research.

Landis and Koch (1977, p. 166) proposed several labels to describe the relative strength of the agreement derived by κ statistics. The strength of the agreement ranges from poor (< 0.0), slight (0.00-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), to almost perfect (0.81-1.00). They noted that their categorisation is arbitrary to a degree, and has to be evaluated carefully for specific research context.

Fleiss et al. (2003, p. 604) also provided recommendations for acceptable levels of reliability for the κ statistics based on Landis and Koch, ranging from poor (below 0.4), fair to good (0.40-0.75), to excellent agreement (above 0.75). These recommendations are based on the experience of these researchers stemming from their research on reliability.

Stemler and Tsai (2008, p. 48) provided guidelines for acceptable inter-coder reliability values for exploratory research studies. Acceptable values are 70% per cent agreement, Cohen's κ of 0.5, Pearson's r of 0.7, and Cronbach's α of 0.7, and an ICC of 0.6.

After outlining the measures used to describe the quality of content coding, the focus is now on the metrics frequently used to describe the performance of machine learning algorithms, which is important when evaluating the performance of the automated reflection detectors.

The following table combines several terminologies from content analysis, machine learning, and statistics in general.

	Relevant/ Actual positive/ H_0 is false (in reality H_1 is true)	Irrelevant/ Actual negative/ H_0 is true
Retrieved	Correct inclusions	Commissions
Predicted/test positive	True positive (TP)	False positive (FP)
Reject H_0 (decision to accept H_1)	Right decision	Type I error (α-error)
Not retrieved	Omissions	Correct exclusions
Predicted/test negative	False negative (FN)	True negative (TN)
Fail to reject H_0 (decision to accept H_0)	Type II error (β-error)	Right decision

Table 4: Combined confusion matrix

Based on Table 4, important metrics such as the accuracy, sensitivity, and specificity are briefly described. These are measures cited frequently in the literature to evaluate machine learning algorithms (for example, see Witten et al. (2011, p. 147-187)).

Accuracy is the sum of all true positives and negatives divided by the sum of all cell values of the confusion matrix:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

Here, accuracy is the same as the per cent agreement outlined above because it calculates the ratio between the sum of all positive and negative 'agreed' cases and all other cell values.

Two other measures that are important for describing the performance of predictive models are sensitivity and specificity. Sensitivity, also referred to as the true positive rate, or recall, is defined as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

Specificity, or the true negative rate, is defined as:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

The receiver operating characteristics (ROC) curve is a tool for inspecting the trade-off between sensitivity and specificity. ROC visualises the sensitivity (true positive rate) against the false positive rate (1-specificity/true negative rate) (see also Bradley (1997) for a discussion on ROC curves). ROC can be summarised in a metric known as the area under the ROC curve (AUC). AUC of 0.5 indicates random performance without predictive value, whereas a value of 1.0 indicates a perfect classifier.

Several other measures for evaluating the quality of machine learning algorithms exist (e.g., Powers (2011)). Among them is the positive predicted value (also referred to as precision) defined as $\frac{\text{TP}}{\text{TP}+\text{FP}}$, negative predicted value ($\frac{\text{TN}}{\text{FN}+\text{TN}}$), and F_1 value defined as $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$).

From the discussion on these metrics, it can be inferred that there is not a single best performance measure. This suggests that it is best to report several measures to evaluate the performance of the machine learning algorithms. In order to cover most of the measurements reported in Section 3.1.4 'Manual reflection detection performance' and Section 3.2.4 'Automated methods performance', the following measures are used in the evaluation chapter: per cent agreement/accuracy because it is the most frequently reported metric in Section 3.1.4 'Manual reflection detection performance'; Cohen's κ because it is mentioned both in the related manual methods and automated methods sections; ICC and Krippendorff's α in order to cover most of the measures stated in Section 3.1.4 'Manual reflection detection performance'; and Gwet's AC_1, although it has not been cited in Section 3.1 'Manual methods to detect reflection' or Section 3.2 'Related automated methods', but it might be useful to include in case other researchers want to compare their results. Cohen's κ, Krippendorff's α, Gwet's AC_1, and ICC are all measures of inter-rater reliability.

Although Cronbach's α is mentioned in Section 3.1.4 'Manual reflection detection performance', it is not included because it is a measure of internal item consistency, and not of inter-rater reliability (see above).

In addition to these main measures, in order to answer the research question, the following measures are added to the discourse: specificity (true negative rate), sensitivity (true positive rate or recall), and AUC.

All measurements are calculated with the R environment for statistical computing and graphics (R Core Team, 2014). Gamer et al. (2012) provided the implementation for Cohen's κ and Krippendorff's α, and Revelle (2013) the implementation for ICCs. The implementation of Gwet's AC_1 (Gwet, 2012) can be found online[1]. AUC and the ROC curves are generated with the pROC[2] package developed by Robin et al. (2011).

1 http://www.agreestat.com/r_functions.html
2 R package pROC: Display and analysis of ROC curves. Version 1.7.3. http://cran.r-project.org/web/packages/pROC/index.html.

4.3 UNIT OF ANALYSIS

The analysis unit results from the research goal, and it is determined by the researcher. A unit is everything that has a distinct meaning to an analyst (Krippendorff, 2012, p. 98-111,277). Common units in text analysis are words, clauses, sentences, paragraphs, and the entire document.

There are two general guidelines to remember when defining the analysis unit (Krippendorff, 2012, p. 102). First, they have to be sufficiently large to capture the meaning of the unit. If the text segment is too small, any meaning that could be derived from the context of the segment might be lost. If the segment is sufficiently large, the meaning is less likely to be missed, adding to the validity of the study. Second, smaller units tend to be more reliable because fewer contexts have to be considered, which might lead to more unanimous interpretation. More contexts might lead to more and different interpretations.

The automated methods outlined in Section 3.2.3 'Machine learning approaches' mentioned as unit of analysis either sentences, text segments, or the entire text (messages, blog post, etc.). Table 3 of Section 3.1.4 'Manual reflection detection performance' listed the units used by the coders of the reflective writings. The analysis unit ranges from the entire text to sentence parts. Usually, the entire text is used as an analysis unit when determining the reflection levels, and smaller units for the descriptive reflection elements. Overall, this suggests that the analysis unit depends on the aim of the study. A more coarse-grained analysis unit is used if the goal is to make statements for each single text, and a more fine-grained analysis unit is used if the aim is to analyse the occurrences of reflection breadth categories in texts. This research on the manual content analysis of reflective writing suggests that

smaller units (paragraph, sentence, etc.) are more suitable for descriptive reflection models.

4.4 SAMPLING

Sampling is an important technique that helps make informed decisions on the generalisability of research findings. The focus of this section is on statistical generalisation.

The general question is to what extent does a smaller sample size represent a larger one with the aim of representing the entire population/universe of texts (Krippendorff, 2012, p. 112). This is about choosing a representative subset for a given universe. The assumption is that by finding a representative subsample, the conclusions drawn are the same for the entire population. Sampling plans can help determine appropriate sample sizes and choices.

Ideally, a sample is drawn randomly from the entire population. However, in order to do this, the population of all these texts must be known. If the sample is sufficiently large, the results allow representative and valid inferences on the entire universe.

Rather than considering the entire universe of texts, a sample frame ample for a particular research question can be chosen that consists of texts available to the researcher. From this sample frame, the actual sample is chosen. The sample frame should ideally represent the universe.

However, the method of drawing samples can introduce bias that can negatively affect the representativeness. Several sampling techniques relevant for the automated reflection analysis are presented and discussed according to their representativeness.

Drawing subsets from the universe can be performed with either probability or non-probability methods (Riffe et al., 2005, p. 97ff.).

Some examples of sampling methods that consider probability are simple, systematic, and stratified random (Krippendorff, 2012, p. 114 ff.). The requirement for probability samplings is a complete list of units of the sample frame. For simple random sampling, several units are selected randomly. Systematic random sampling means that from a randomly chosen unit, every other xth unit is chosen. However, if the sample is ordered in a way that conflicts with the decision of the xth unit, it can introduce bias. For example, this applies to events that are time-dependent. Let us assume that a researcher attempts to estimate course participation in an online forum, and only considers every 7th day (e.g., Sunday); this might not reflect general course participation. Stratified sampling means to group the universe into several strata. From every stratum, the sample is determined either by simple random or systematic random sampling. For example, a text corpus consists of texts. Each text belongs to a genre. The frequency of texts that belong to a genre might vary. For each genre or stratum, a sub-sample can be drawn randomly based on the frequency of texts available in each stratum.

Methods not based on probability are multistage, convenience, snowball, and relevance sampling (Krippendorff, 2012, p. 114 ff.). Convenience sampling is a method through which the researcher includes or excludes units based on their convenience. There is no effort made to sample a population. Snowball sampling is a technique that systematically includes or excludes texts based on an initial set of units. The extension of the sample is based on a sampling criterion. Relevance sampling or purposive sampling is a technique that retrieves units relevant for answering the research question. The goal is to systematically reduce documents to a manageable amount based on the relevance criteria.

With the assumption that reflections are relatively rare (see Ullmann et al. (2013) for empirical evidence that supports this assumption), there are fewer texts that contain reflections than texts that do not express a reflective stance. This introduces certain

trade-offs to be considered when sampling texts with respect to the generalisability of the sample. Random sampling from large text corpora (for example, the British national corpus[3]) results in a text collection, which only contains a small proportion of texts that contain reflections.

A small set of texts that contain the sought characteristic traits of reflection is diametrical to the requirements of automated machine learning methods that usually require a larger set of training data. With this outlined, the trade-off is between reaching high generalisability using a large representative sample with the cost of many texts with few reflections, and custom samples with lower generalisability, but with higher amount of reflections.

After the discourse on evaluation criteria, measurements, choice of analysis unit, and sampling strategies, the next important decision to discuss is the selection of the machine learning algorithms.

4.5 MACHINE LEARNING ALGORITHMS

There is a large variety of machine learning algorithms from which to choose. For example, Fernández-Delgado et al. (2014) listed 179 machine learning algorithms for only one area of machine learning, namely, classification problems. Whereas some perform well on many different datasets, the question is, do machine learning algorithms perform well in the context of the detection of reflection?

When faced with the problem of choosing good candidate machine learning algorithms, several considerations can be made.

In general, machine learning can be distinguished as either a supervised or unsupervised learning problem (James et al., 2013, p. 26 ff.). Supervised methods are based on a defined variable that the machine learning algorithms seek to predict

3 http://www.natcorp.ox.ac.uk/

based on several predictor variables. For example, an analysis unit is labelled as reflective or descriptive. The unit text serves as label predictor. Unsupervised methods work without this response variable, and their aim is to detect structure in the existing data without the supervision of a response variable. Some examples are clustering or the discovery of topics.

The immediate problem is a supervised problem because the aim of this thesis is to predict labels from text inputs relevant for reflection.

Another distinction can be made based on the type of obtainable variables. The data can be either quantitative (numerical) or qualitative (categorical) (James et al., 2013, p. 28 f.). An example of quantitative data is stock market prediction based on historical stock market prices. In this context, qualitative variables are, for example, the categories of the common reflection categories, or reflective vs. descriptive. This distinction can help determine the type of model to use: a model more suitable for regression problems (quantitative), or a model more suitable for classification problems (qualitative). However, it is not strictly an either/or decision because some machine learning algorithms can work on both problem spaces (James et al., 2013, p. 28).

For the context of this thesis, the variable to predict is qualitative/categorical, and not quantitative. Therefore, the problem space is one of classification. In addition, the variables that predict the outcome variable (e.g., reflective or descriptive) are also qualitative. For example, they might be all words of an analysis unit.

With these decisions, the choice of algorithms is reduced to machine learning algorithms for classification problems. There is a wide variety of machine learning classifiers available (for recent reviews, see Caruana and Niculescu-Mizil (2006); Kotsiantis et al. (2007); Kotsiantis (2007); Khan et al. (2010); Aggarwal and Zhai (2012); Fernández-Delgado et al. (2014); and Aphinyanaphongs et al. (2014)). The following paragraphs provide the rationale for the selection process of classifiers that are good candidates for reflection detection.

According to Aggarwal and Zhai (2012, p. 165 f.), some of the key methods used for text classification are decision trees, rule-based, Support Vector Machine (SVM), Neural Network, and Naïve Bayes classifiers. All, with the exception of the rule-based algorithms, were applied to the research outlined in Section 3.2.3 'Machine learning approaches'. These are good candidates because they were successfully applied to related concepts of reflection.

The first two, decision trees and rule-base classifiers, have the benefit that their models can be interpreted intuitively because they either build a tree structure – a decision tree, or they generate rules that follow the common premise and conclusion pattern. They do not belong to the highest performing classifiers (Fernández-Delgado et al., 2014), but they do have the potential of fostering the understanding of reflective writing because they exhibit patterns in a more easily interpretable way.

In addition to the classifiers outlined in the related automated methods section, Random Forests is included to the set of classifiers for the final evaluation because it shows constantly good results on an array of different datasets in the study of Fernández-Delgado et al. (2014).

The candidate classifiers for this thesis are tree-based, rule-based, SVM, Neural Network, Naïve Bayes, and Random Forests. These are used to test the research question of the thesis.

In order to structure the argument used to answer the research question, the candidate classifiers are summarised into the three categories listed below. Each of these three categories are related to one of the three lines of investigations described in Section 1.1 'Research questions'.

Tree-based models: tree-based models form a decision tree of the data. The tree can then be used to determine whether a particular sentence is reflective. As a simplified example, let us consider a tree that consists of three nodes (one root and two terminal nodes). The root node checks whether a sentence contains the token 'I'. If so, the

sentence is classified as yes (terminal node 'yes'); otherwise, it is classified as 'no'. The benefit of tree-based models is that they can be highly interpretable. One of the limitations is that they may not produce the best performance. Tree-based models are used to investigate the first line of investigation, 'I1: Can tree-based machine learning algorithms detect the difference between descriptive and reflective texts segments?'

Rule-based models: with rule based models, the classification decision is defined in one or more rules that follow the form of if-then statements. Only if the condition is true, is the conclusion true. For example, if the sentence contains the term 'I', it is a reflective sentence. As with the tree-based models, rule-based models are highly interpretable, but they may be lacking in predictive performance. Rule-based models are used to investigate the second line of investigation, 'I2: Can rule-based machine learning algorithms detect the difference between descriptive and reflective text segments?'

High performance models can be discerned from tree-based and rule-based models because of their complexity in prediction capability. They usually have a higher predictive capability than the other two types, but their models may not be easy to interpret. The following classifiers are selected as candidates to investigate the third line of investigation 'I3: Can high performance machine learning algorithms detect the difference between descriptive and reflective text segments?': SVM, Neural Network, Naïve Bayes, and Random Forests.

The first two types of models are interesting because of their ability to generate decision trees and rules that can be seen as frequently co-occurring word tokens. Such models find patterns in the data that can be helpful in obtaining a better understanding on the nature of reflection expressed in texts.

The third type uses more complex mechanisms to automatically classify sentences. They produce models that usually perform very well, but are generally not easy to interpret. The research focus with this third type is less on what constitutes reflection,

and more on the exploration of the predictive capabilities of these algorithms in the context of automated analysis of reflection.

Investigations of all three types of machine learning algorithms help obtain an understanding on the quality of the algorithms for automated analysis of reflection.

After the discourse on the selection of good candidate machine learning classifiers for the problem of reflection detection, the following sections list the specific classifiers used in the evaluation part of this thesis. In addition, these sections provide information on the software that implemented each of the classifiers.

4.5.1 *Tree-based models*

Several tree-based models are explored, all of which build a tree structure of the data. The models can be differentiated in the way they find the optimal division of a node into sub-nodes, and in how the trees are pruned to achieve better fit for unseen data.

The following algorithms for constructing tree structures suitable for classifying text are considered:

CART trees (Classification and regression trees). Trees are built in two stages: the first builds the tree, and the second prunes the tree to avoid overly complex trees that do not generalise well. The tree building stage starts by finding the variable that best separates the tree into two subsets. Then, for each of the two subsets, the best division is determined. This process continues until a stop criterion stops the growth of the tree. In the second stage, the full-grown tree is pruned to the right size using cross-validation to alleviate over-fitting. An over-fitted model predicts every instance of one dataset very well, but does not predict instances of another datasets very well. It has learned the noise of one dataset that reduces its performance for other datasets. The right choice of model parameters helps to avoid over-fitting. Here, two variants for finding the right size of the tree are available. The first parameter is based on a

complexity parameter (cp). Trees tend to be smaller with bigger complexity parameter values. The second parameter controls the maximum depth of a tree.

The implementation used here is from the R-package rpart[4] (Recursive Partitioning and Regression Trees) created by Therneau et al. (2014). The authors of the package implemented functions based on the book by Breiman et al. (1984).

Conditional inference trees. This algorithm uses a conditional inference framework (Hothorn et al., 2006) for splitting the tree. The data are evaluated with the ctree algorithm of the R-package party[5]. The ctree algorithm is tuned once for minicriterion (this is the value of the test statistic that regulates a separation), and once for maxdepth (maximum tree depth).

C5.0 Decision Trees. This algorithms was proposed by Quinlan (1993). The implementation used here can be found in the C50[6] package by Kuhn et al. (2014a).

C4.5-like trees. This algorithm is an older version of the C5.0 algorithm by Quinlan (1993) (see above). The implementation used here is the J48 algorithm described by Witten and Frank (2005). J48 is a C4.5-like implementation, and it can be found in the R package RWeka[7] provided by Hornik et al. (2009). The algorithm has the confidence threshold C as the tuning parameter.

4.5.2 *Rule-based models*

Rules have the benefit of forming if-then statements that are highly interpretable. Rule-based systems can be either built by experts who manually generate rules for a knowledge-based system, or the rules can be derived from the data.

For example, Ullmann et al. (2012) explored the potential of manual reflection detection rules. A set of rules was developed aimed at detecting those sentences

4 Version 4.1-8. http://cran.r-project.org/web/packages/rpart/index.html.
5 Version 1.0-15. http://cran.r-project.org/web/packages/party/index.html
6 Version 0.1.0-19. http://cran.r-project.org/web/packages/C50/index.html.
7 Version 0.4-23. http://cran.r-project.org/web/packages/RWeka/index.html.

relevant for reflection. The rules built on each other and on the last level of aggregation are indicative of whether a text is reflective.

Here, the rules are not created manually. Instead, the following algorithms are chosen to form rules based on the underlying data. The following rule-based models were chosen in order to test their suitability for the detection of reflection.

C5.0 Rule-Based Models. The C5.0 group of algorithms has a version for decision trees (see Section 4.5.1), but also a rule-based version. The R-package C50[8] by Kuhn et al. (2014a) is used to test the model performance.

PART generates rules from partial decision trees (Frank and Witten, 1998). This algorithm can be tuned over a threshold parameter.

OneR generates one rule based on the variable with the lowest error rate (Witten and Frank, 2005). This is an extremely simple model because it only produces a single rule. This algorithm is included in order to find a baseline of the performance of the machine learning algorithms.

JRip is a Ripper algorithm (repeated incremental pruning to produce error reduction) for finding rules (Witten and Frank, 2005). The numopt tuning parameter is the number of optimisations.

The implementations used for PART, OneR, and JRip can be found in the R package RWeka[9] developed by Hornik et al. (2009).

All the algorithms Section 4.5.1, Section 4.5.2, and Section 4.5.3 can be executed with the R-package caret[10] created by Kuhn et al. (2014b). This is a classification and regression training framework that unifies and standardises the execution of machine learning algorithms.

8 Version 0.1.0-19. http://cran.r-project.org/web/packages/C50/index.html.
9 Version 0.4-23. http://cran.r-project.org/web/packages/RWeka/index.html.
10 Version 6.0-30. http://cran.r-project.org/web/packages/caret/index.html.

4.5.3 *High performance models*

SVM. This is a class of models aimed at finding the optimal hyperplane to separate classes. By adjusting their kernel function, SVMs can be extended to capture patterns that are not linear. The implementation used here is from the R package kernlab – the Kernel-based Machine Learning Lab[11] developed by Karatzoglou et al. (2004).

Three kernel functions are explored: the linear kernel tuned over several cost candidates; the radial kernel tuned over sigma and cost; and the polynomial kernel tuned over degree, scale, and cost. All three kernels have been successfully applied to text classifications tasks (Joachims, 1998; Hornik et al., 2006).

The cost parameter controls the smoothness of the function (higher cost generates less smooth functions). Sigma controls the kernel width and it can be estimated automatically by the package. Degree and scale are the parameters of the polynomial kernel function.

Neural Networks are techniques inspired by the mechanisms of the brain. A Neural Network is a network that starts at the input nodes (neurons) of the predictor variables, and ends in an outcome node with the prediction. In the middle, there is a network of hidden nodes. This layer is where the network learns the function to map the input nodes to the output nodes.

The chosen implementation is the R package nnet-Feed-forward Neural Networks and Multinomial Log-Linear Models[12] developed by Venables and Ripley (2002). The nnet algorithm can be tuned over several candidate parameters, namely, size and decay.

Random Forests are algorithms similar to the decision tree models used in Section 4.5.1. In fact, it is an ensemble of decision trees. The algorithm generates from selecting randomly a subset of predictors decision trees over several iterations. The

11 Version 0.9-19. http://cran.r-project.org/web/packages/kernlab/index.html.
12 Version 7.3-8. http://cran.r-project.org/web/packages/nnet/index.html.

result of this is many different trees - a forest of trees. To classify an input, the algorithm runs through all decision trees. The results from all these decisions are aggregated and reported as the classification output.

The implementation used is the R package randomForest – Breiman and Cutler's Random Forests for classification and regression[13] developed by Liaw and Wiener (2002) based on the code from Breiman (2001). The Random Forests implementation can be tuned over several mtry candidates, which is the amount of randomly selected input features.

Naïve Bayes is an algorithm based on Bayes' rule. It is naïve because it assumes that all predictors are independent from each other. This has the benefit of quick processing while producing good performance in many cases.

The implementation used is one from the R package klaR - Classification and visualisation[14] - developed by Weihs et al. (2005) by extending the code from Meyer et al. (2014). The tuning parameters are fL that control the Laplace correction, and usekernel, which is Boolean. If set to true, a kernel density estimate is used for density estimation; if set to false, the normal density is estimated.

This concludes the discussion on the selection of promising machine learning algorithms for the detection of reflection in texts. Based on several decision criteria, a list of candidate classifiers was compiled. These are seen as strong candidates for the problem of reflection detection because they have been applied successfully to related concepts of reflection as outlined in Section 3.2.3 'Machine learning approaches', and further, these candidate classifiers show generally good performance when applied to a variety of datasets. In addition, high performing classifiers and classifiers that can generate interpretable results were chosen. The latter can potentially help understand better the structure of reflective writings.

13 Version 4.6-7. http://cran.r-project.org/web/packages/randomForest/index.html.
14 Version 0.6-11. http://cran.r-project.org/web/packages/klaR/index.html.

After outlining general methodological considerations in Section 4.1 'General methodological considerations'; the importance of the evaluation criteria on reliability, validity, and objectivity; several measurements (see Section 4.2 'Evaluation criteria and metrics'); the discourse on the choice of the analysis unit (see Section 4.3 'Unit of analysis'); sampling strategies (see Section 4.4 'Sampling'); and the discussion on the selection of suitable machine learning algorithms (see Section 4.5 'Machine learning algorithms'), the next section describes the research design used to answer the research question.

4.6 OVERVIEW OF RESEARCH DESIGN

This section explains the research design used to answer the research question. The research design is based on the previously outlined methodological considerations. There are two parts to discuss: first, the process applied to generate the dataset on which the machine learning algorithms are trained and tested, and second, the design used to evaluate the performance of the machine learning algorithms.

Figure 1 shows a high-level overview of the research design with the data generation process at the top and the evaluation at the bottom.

The data generation step (upper half) describes the process to generate the dataset used in the evaluation step. It starts with a text collection. The text collection is then annotated according to indicators (also called operationalisations or proxies) of the common reflection categories. The result of the annotation task is a collection of annotated units that serve as input to the evaluation step. The evaluation step (lower half) uses the dataset of text units as input. This dataset is divided into two parts: the training data used to generate the machine learning models, and the test data used to assess the models generated in the training step. The assessment of the models is the

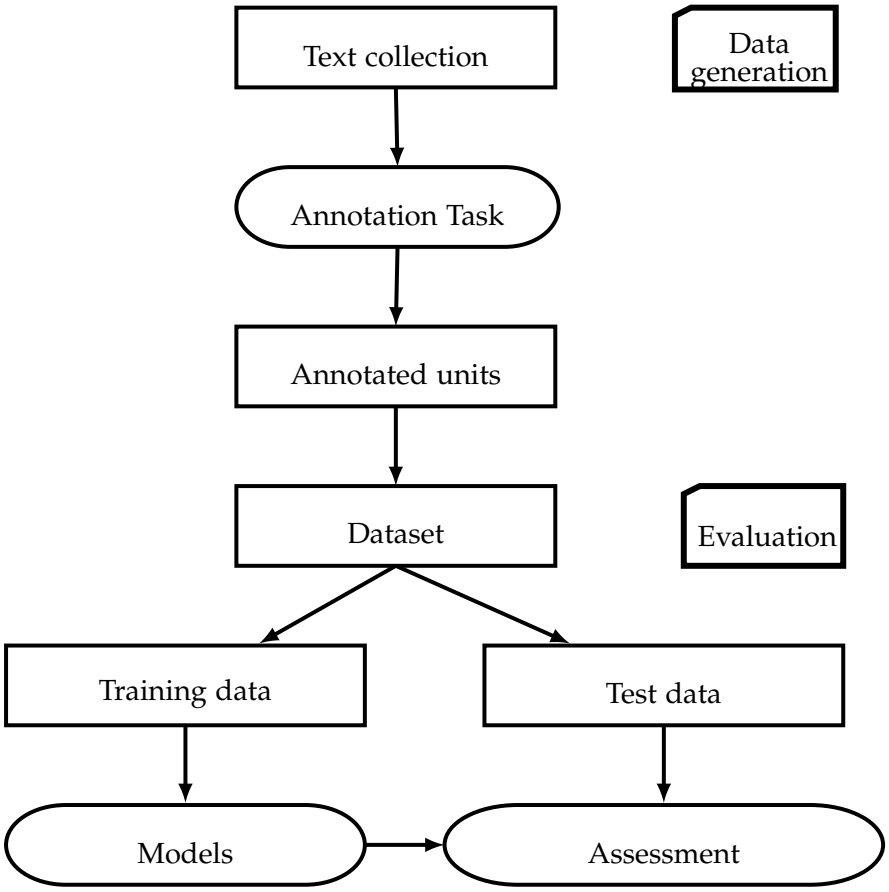

Figure 1: Overview of research design

final step, and it provides the performance estimates necessary to answer the research question.

The next two sections describe the details of this high-level overview of the research design. It starts with the description of the process applied to generate the dataset.

4.6.1 *Dataset generation process*

The review of the automated methods in Section 3.2.3 'Machine learning approaches' suggested that large datasets are necessary to apply machine learning to related constructs of reflection. McKlin (2004) reported 1,180 messages for training and 295 for testing. Kovanovic et al. (2014) based their study on 1,747 messages, and Dönmez et al. (2005) on 1,255 text segments. Considering that reflection occurs relatively rarely

in texts (Ullmann et al., 2013), intensified the requirements for a large dataset for a sufficient amount of reflective training instances. This insight is one of the main driving forces behind the process for the generation of the datasets.

A dataset had to be generated, because to our best knowledge, there is no public dataset available that contains labelled units according to the categories of reflective writing.

The start of the dataset generation process is a collection of texts, and the output is a dataset that contains annotated text units. Each step is now described. Figure 2 shows the schematic representation of the data generation process.

The process starts with the **identification of suitable text collections** that should first contain reflective writings. Second, such collection should be of a reasonable size, and third, it should be from a context similar to the research on the manual content analysis of reflective writing.

The size of the text collection is important. If the text collection is too small, there is not sufficient data to train the machine learning models. In addition, it should be considered that texts do not contain an abundance of reflection (Ullmann et al., 2013). On the assumption that 10% of the units of a suitable reflective text collection are reflective, and considering the aim of reaching approximately 500 reflective units (the primary concern is reflective units because descriptive or non-reflective units are more frequent), approximately 5,000 units are necessary, thus underlying the necessity of a large text collection.

Further, the research on the manual analysis of written reflection is mostly conducted in an academic context. Mostly student writings of the participants of courses or trainings were analysed. It would be beneficial for the text collection to stem from students writings to strengthen the comparison of the automated method with the manual method.

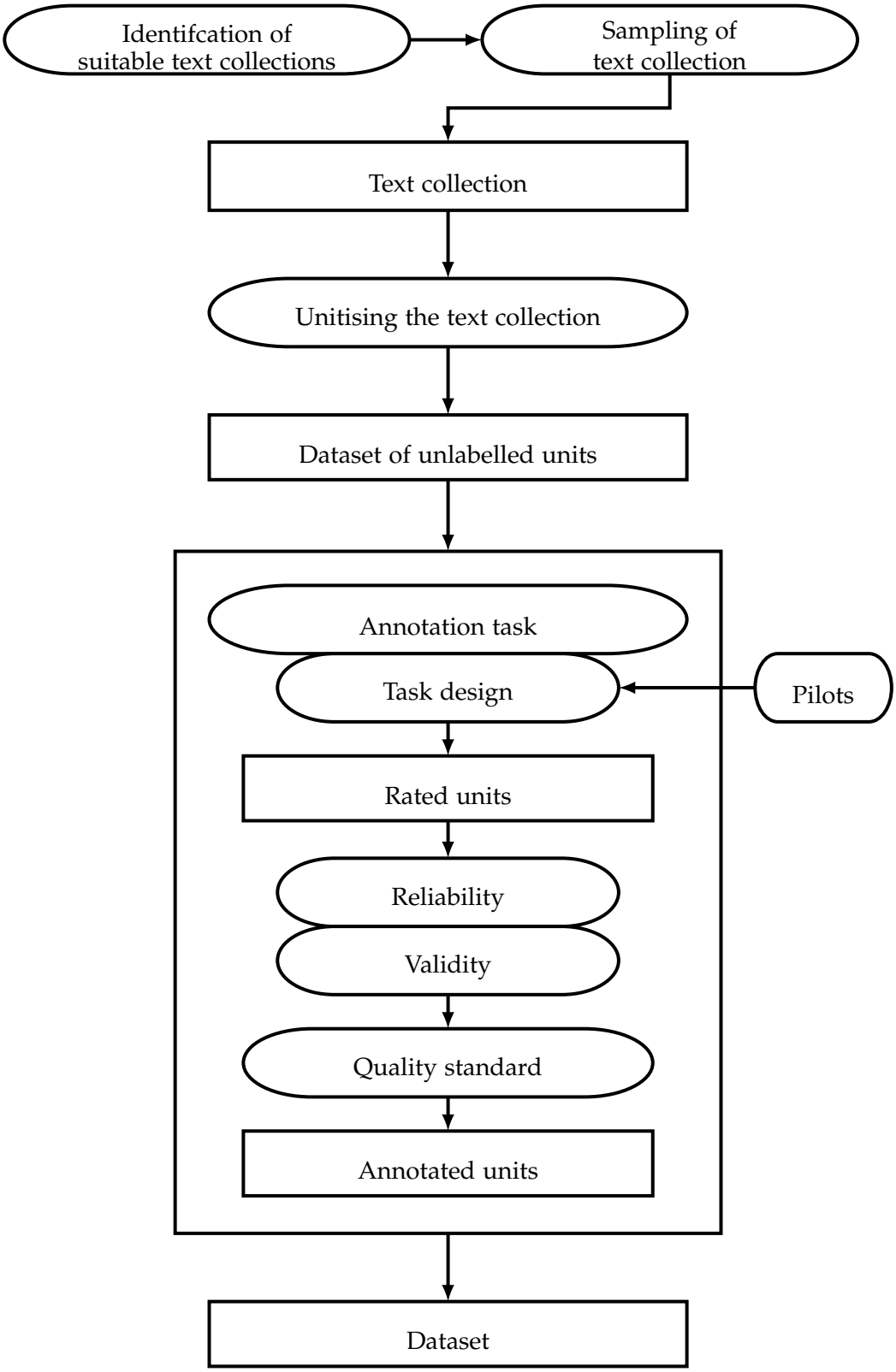

Figure 2: Overview of data generation process

Sampling of the text collection: Several sampling techniques have been described Section 4.4 'Sampling'. Instead of annotating the whole text collection a representative sample suitable for the research project can be drawn from the text collection. The

chosen text collection consists of approximately 2,800 student assignments. Close inspection of this corpus reveals that only a subset of the corpus contains the students' personal accounts. Most assignments are written in a factual style, which is common for academic assignments. In order to maximise the probability of texts that contain reflection, a form of sampling has to be chosen. Considering the generalisability of the results, the first choice is a probability-based strategy similar to the random sampling technique. Here, however, a random sample technique would have included many texts not relevant for the aims of this research, because they do not contain reflection. This excludes random sampling as a strategy. From the non-probabilistic sampling strategies discussed in Section 4.4 'Sampling', the relevance sampling technique is chosen. It is a purposive technique that requires the text collection to be sampled according to explicit relevance criteria. It is a technique that retrieves units relevant for answering the research question. Most of the texts are included based on criteria that can be repeated by other researchers, because they make use of text features that can be implemented in a simple algorithm. The implementation details of this triangulation are discussed in Section 5.2 'Sampling text collection'. The outcome of this step is a **collection of texts**.

Unitising: after determining the final collection of texts, the next decision is the analysis unit. As outlined in Section 4.3 'Unit of analysis', text-segments, compared with entire texts, are more suitable for the analysis of the descriptive (breadth) categories of reflective writing. The choice for the analysis unit is sentences. Single sentences as units were already successfully applied in the research on the manual analysis of reflective writings (see Table 3). In addition, texts can be divided automatically into sentences. A program has been written to divide each text into sentences. This process has the benefit that the analysis units can be determined in an objective way (see the evaluation criterion of objectivity in Section 4.2 'Evaluation criteria and metrics'). The result of this process is a **dataset of unlabelled sentences**.

Several **pilots** preceded the actual annotation task in order to determine a suitable coding process. Considering the large amount of annotated sentences necessary for machine learning, the decision was made to use crowdsourcing to annotate the analysis units (see Section 5.5 'Background on crowdsourced text annotation'). Research on crowdsourcing text annotations tasks indicated promising results with regard to the reliability of large scale text annotation tasks (see Section 5.5.1 'Research on crowdsourced annotation quality'). Crowdsourcing is a suitable method for large annotation tasks. The text collection used in here is large in size. The final text collection contains more than 5,000 sentences, which is equivalent to approximately 130,000 words (the text collection has a higher word count than this thesis). As outlined in Section 5.5 'Background on crowdsourced text annotation', the crowdsourcing provider distributes small tasks to many people who work in parallel on the task. With this process, each sentence is rated by many coders, rather than relying on a small number of coders as is predominant in the traditional content coding setting (Section 3.1.4 'Manual reflection detection performance'). Several pilots with smaller datasets were conducted in order to develop a process that yielded sufficiently high quality labels for each sentence (see Section 5.6 'Summary of pilots'). The experiences observed during the pilots informed the **task design** used for the annotation task.

Annotation task: in the annotation task, the sentences in the text collection are rated. Based on the ratings, the sentences are annotated with regard to the categories for reflection detection. A **task design** is created that ensures that the entire rating process is conducted under the same conditions. The task design specifies the coding instructions, coder training with test questions, and the coding schema for the sentences. The rating schema is based on the common reflection categories (Section 2.3.2 'Common reflection categories'). The task design is based on the experiences observed during the pilots (see Section 5.6 'Summary of pilots'). The task

design is described in Section 5.7.2 'Task design'. The task was then administered to the coders. All coder ratings were collected, and the result is a collection of **rated units**. Based on the ratings of all sentences, the **reliability** of the coding process is determined (see Section 5.7.4 'Reliability'), and the **validity** of the indicators is evaluated (see Section 5.7.5 'Validity').

Quality standard: the next step controls the quality of the annotations for the final dataset by applying a standard. The quality of the dataset is important because the machine learning algorithms build their models from these examples; therefore, the dataset should only contain units that represent the construct. Consequently, a unit should only be associated with an annotation when there is sufficient evidence that the sentence indeed expresses this annotation. A strict standard is set based on the criterion of how much support each sentence receives from all coders. The standard is set to 80%, which means that at least 80% of the raters (a four-fifth majority) have to agree on a particular sentence in order to label it. For example, a sentence is labelled as reflective if at least eight out of ten coders rate the sentence as reflective. A more lenient approach is to use simple majority voting (>50%). A four-fifth majority, compared with a simple majority, is a stricter standard for labelling a sentence because it demands more evidence for a label. This ensures that the labelled sentences entered into the dataset are of higher quality compared with a simple majority vote approach. It also strengthens the validity of the dataset because only sentences that received substantial support of being reflective enter it. The result of this process is a collection of **annotated units**. The annotated units are organised in datasets.

Dataset: the outcome of the data generation process is a dataset of annotated sentences. The labels are based on the chosen indicators of the common reflection categories (see Section 2.3.2 'Common reflection categories'). The outlined data generation process is the blueprint for Chapter 5 'DATASET GENERATION' that

applies this process. The outcome of the data generation process is the datasets that serve as input for the research design.

4.6.2 *Research design*

The research design for the evaluation of the machine learning algorithms on the problem of reflection starts with the dataset of labelled sentences, and ends with the calculated performance measures in the model assessment step. Figure 3 illustrates all the steps of the research design.

The research design consists of data preparation and machine learning steps. The overarching aim of this research design is to develop a setup that is the same for all the machine learning algorithms described in Section 4.5 'Machine learning algorithms'. This strategy ensures that the conditions are the same for all machine learning algorithms, and thus their differences can be attributed to their functioning and not to context factors.

In the **data pre-processing** step, the dataset is transformed into a format that can be used by the machine learning algorithms. The following paragraphs discuss the two steps: feature construction and selection [15].

Feature construction: A feature construction choice is to transform text into single words (unigrams) (Sebastiani, 2002, p. 10). The text is transformed into a multiset. In this model, the ordering of words does not play any role. For example, the multiset representation of the sentence 'rose is a rose' is {rose,rose,is,a}, where the word 'rose' occurs twice, and 'is' and 'a' occur only one time. This can be represented as vector [2,1,1], where the three types of words ('rose', 'is', and 'a') are mapped to their frequency.

15 Feature selection can also be part of the machine learning phase (see Mladenic (2011)).

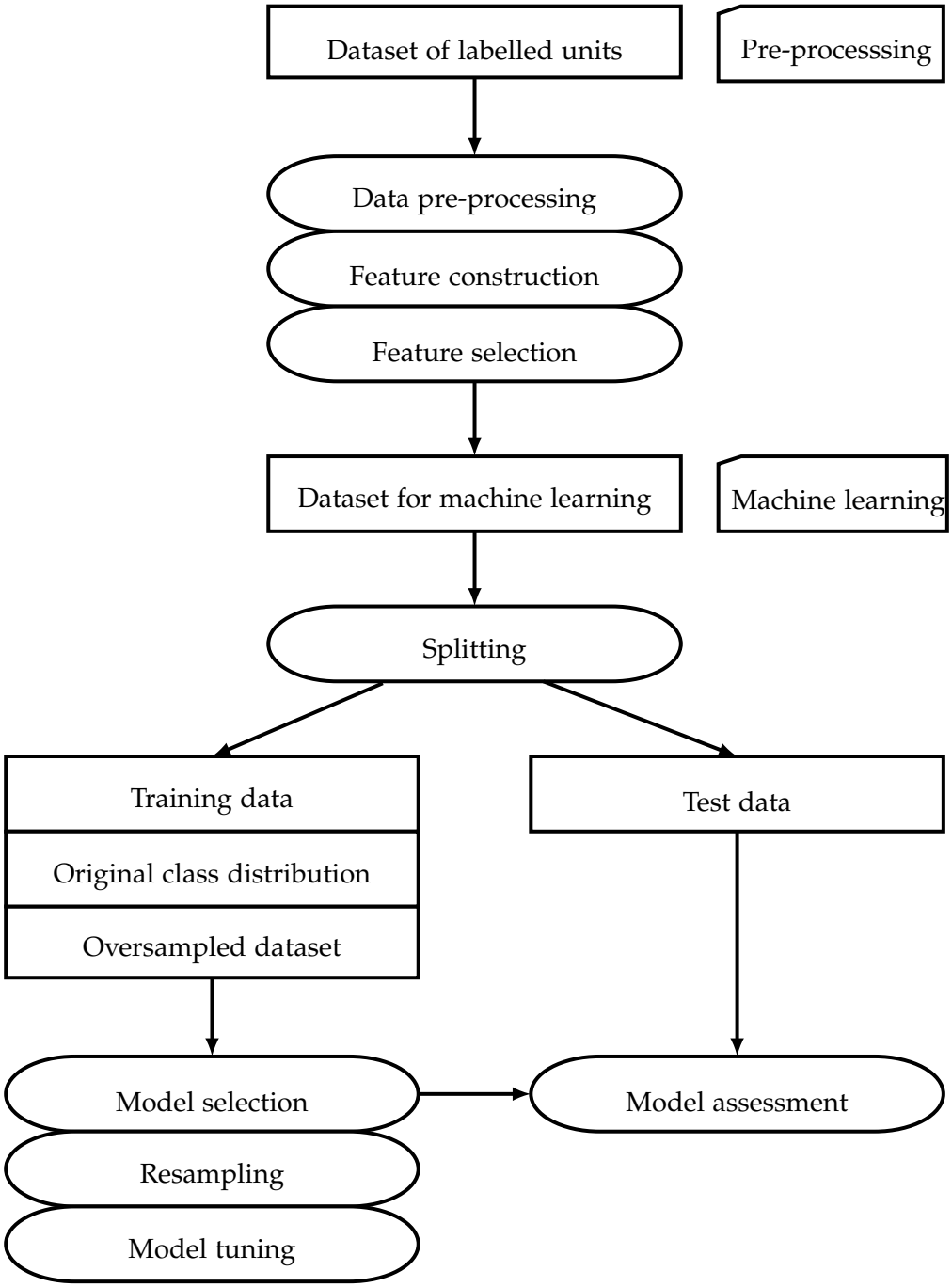

Figure 3: Overview of research design

The approach taken here is to covert the sentences to unigrams, and instead of a multiset representation, a set representation is used. This means that if a word such as 'rose' occurs twice in a sentence, it is counted only once (in our example, the count vector is [1,1,1], instead of [2,1,1]). Unigrams are derived based on word boundaries. This is a more simple representation than the multiset representation. It de-emphasises the importance of repeated words.

In addition, all unigrams are converted to lower case and stemmed using the stemming algorithm from Porter.

There are several other methods for constructing features from text (see Brank et al. (2011); Blake (2011)). In the context of this thesis, Rosé et al. (2008) and Kovanovic et al. (2014) tested a variety of feature types on the performance of the classifier (see Section 3.2.3 'Machine learning approaches'). Their research indicated slightly better performance with other types of features than unigrams.

Feature selection: The transformation of texts into features usually produces a large feature space. Feature selection is used to reduce such feature space. The aim is to remove all those features that do not provide information for the classification, or to remove those features that introduced noise to the model (Manning et al., 2008, p. 271).

There are many methods for selecting features (Forman, 2003; Mladenic, 2011). For example, McKlin (2004), used a dictionary to reduce the feature space. Only those words that were part of the dictionary were conserved (see Section 3.2.3 'Machine learning approaches').

Here, the removed features were punctuations, numbers, and white space. In addition, all the features that occurred only ten times or less in the entire dataset were removed. The last step had the greatest impact on the feature space: it ensured that the feature space was not filled with rare unigrams.

These pre-processing steps are applied to all individual datasets, each representing a single operationalisation of the common reflection category. These two pre-processing steps are conducted with the R Text Mining Package tm[16] (Feinerer et al., 2008; Feinerer and Hornik, 2014).

Each dataset now contains feature vectors instead of sentences, along with their labels that represent the absence or presence of a common reflection category. These datasets are the **datasets for the machine learning** task.

16 Version 6.0. http://cran.r-project.org/web/packages/tm/index.html.

Training and test set: Each dataset is then randomly **divided** into training and test datasets. The training dataset is used to train the machine learning models. The test dataset is used to evaluate the performance of the model generated in the training phase. All machine learning algorithms outlined in Section 4.5 'Machine learning algorithms' are trained with the same training set, and the model performance is evaluated with the same test set. The training set consists of 80% of the dataset, and the test set of the remaining 20%. Considering the available data, this ratio is found to be a good compromise for all datasets.

Class imbalance: As outlined in Section 4.4 'Sampling', reflection occurs relatively rarely in texts (see Ullmann et al. (2013)), which can influence the performance of the classifier models. Imbalanced datasets have been identified as a problem of machine learning (Chawla et al., 2004). As we can see later, there is an imbalance in the amount of sentences that are reflective compared with non-reflective/descriptive. Such imbalance is not drastic (see Chawla et al. (2004)); however, it can still influence classifier performance (see also the problem of imbalanced categories in the context of the per cent agreement in Section 4.2 'Evaluation criteria and metrics'). Several remedies for class imbalance were researched at the algorithmic (Chawla et al., 2004) and data levels (Chawla, 2005; Menardi and Torelli, 2012). In order to investigate the potential problems of class imbalance, a data-driven strategies is tested: random oversampling (also referred to as upsampling). Random oversampling generally can improve the results (Japkowicz and Stephen, 2002; Batista et al., 2004). Here, oversampling means that the minority class instances of the training data are repeated randomly to match the amount of the majority class. Undersampling means that the instances are removed randomly from the majority class until both classes are of equal size. The models built with these data are then evaluated on the original imbalanced test set. All models were also trained with the undersampled training dataset. Overall the performed was worse than the original dataset or the oversampled dataset.

Model selection: The aim of the model selection phase is to find good candidate models likely to perform well on the unseen test data. The training data are used to tune the parameters of the models, and to determine the best candidate model using a resampling strategy.

Resampling: Resampling is related to the sampling strategies outlined in Section 4.4 'Sampling'. In machine learning, resampling strategies are used to estimate the performance of a model.

All models are trained with the same resampling strategy. The choice of the resampling technique is repeated k-fold cross-validation (similar to Rosé et al. (2008)). Molinaro et al. (2005) and Kim (2009) indicated that repeated k-fold cross-validation has slight advantages over k-fold cross-validation. Here, the training set is divided randomly into ten folds of approximately the same size. The training model is generated on nine folds, and the tenth fold (the validation set) is used to test the performance. This process continues iteratively for all remaining nine folds. The entire process is repeated five times with a different randomly sampled separation of the data into the ten folds. The performance measure for the 50 validation sets is averaged.

Model tuning: Most models described in Section 4.5 'Machine learning algorithms' have tuning parameters. These are algorithm parameters that cannot be estimated directly from the data. A poor choice of these parameters leads to poor model performance. In order to select good parameter candidates, here, a set of parameters is defined, and for each parameter, a model is generated and evaluated. For each parameter candidate, the average performance over all hold-out samples of the training data (see the point resampling above) is calculated. The model with the highest performance (measured with Cohen's κ) is determined. The parameters of this model are used to generate the final model with all training data. This is the final model used to assess the performance on the test data.

Model assessment: The final model is used to classify all the sentences of the test set. The performance measures outlined in Section 4.2 'Evaluation criteria and metrics' (on page 104) are calculated based on the model predictions and the actual, true values (see Table 4). The ground truth is based on the annotations of the dataset.

The model assessment step concludes the research design. As outlined above, the aim of this research design is to find a standardised approach that can be used over all datasets to assess model performance for reflection and the common categories of reflective writing. The described research design achieves this, thus allowing the comparison of results because all context factors are the same, whereas only the machine learning algorithms vary. This standardisation is at the price of not applying individual strategies that might have yielded better performance. However, the ability to concisely report and compare the results is considered more useful for the aims of this thesis, and outweighs the benefits of maximised strategies for individual machine learning algorithms.

The research design outlined thus far is used on each of the datasets generated with the process described in Section 4.6.1 'Dataset generation process'. The datasets serve as input to the research design. The research design is applied, and the results of the machine learning assessment phase are retrieved. These evaluation results inform the answer of the research questions outlined in Section 1.1 'Research questions'.

Figure 4 shows the concrete instantiation of the research design for both research questions.

As outlined, the overall aim of this thesis is to investigate whether text segments can be analysed automatically using machine learning algorithms to detect the presence (and absence) of key elements of reflection. The first of the two research questions is: 'Q1: Can machine learning algorithms be used to distinguish between descriptive and reflective text segments?' Three lines of investigations are used to answer this research question: 'I1: Can tree-based machine learning algorithms detect

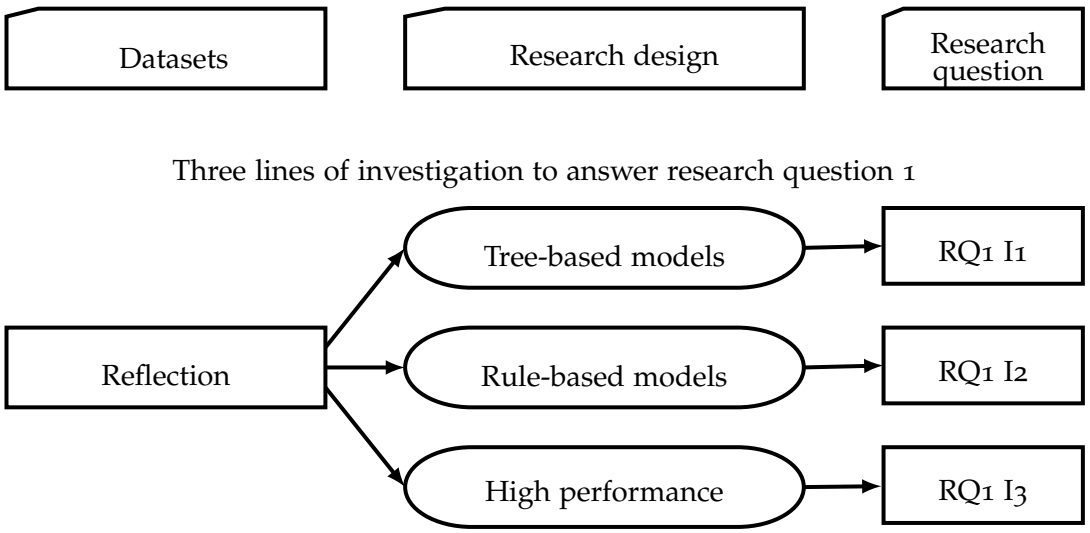

Figure 4: Instantiation of research design

the difference between descriptive and reflective texts segments?', 'I2: Can rule-based machine learning algorithms detect the difference between descriptive and reflective text segments?', and 'I3: Can high performance machine learning algorithms detect the difference between descriptive and reflective text segments?'. As can be seen in Figure 4, each line of investigation has a set of corresponding machine learning

algorithms. Based on the same dataset that consists of reflective and non-reflective sentences, the research design is applied to machine learning classifiers that belong to one of the three groups of tree-base, rule-based, and high performance algorithms (see Section 4.5 'Machine learning algorithms'). The selected tree-based models are listed in Section 4.5.1 Tree-based models', the rule-based models in Section 4.5.2 Rule-based models', and the high performance models in Section 4.5.3 High performance models'. Based on the test set, the performance measures outlined in Section 4.2 'Evaluation criteria and metrics' (on page 104) are calculated for each algorithm. For each line of investigation, a performance profile of the machine learning algorithms is generated to discuss the research question.

In order to answer research question 2, 'Q2: Can machine learning algorithms be used to detect common categories of reflective writing?', several datasets serve as input to the research design. For each common reflection category (see Section 2.3.2 'Common reflection categories'), a dataset is prepared and generated following the dataset generation process outlined in Section 4.6.1 'Dataset generation process'. Each dataset is one concrete indicator (aka operationalisation) of the common reflection categories, which are abbreviated as Exp. (Description of an experience), Feel. (Feelings), Pers. (Personal), Crit. (Critical stance), Persp. (Perspective), and Out. (Outcome). Based on the experiences observed to answer research question 1, a selection of high performing algorithms is chosen. For each dataset, the same models are trained and tested according to the research design process. For each dataset, a performance profile is created that contains the measurements outlined in Section 4.2 'Evaluation criteria and metrics' (on page 104). This profile serves as the evidence to answer research question 2, first for each reflection category, and then in a synopsis for all evaluated operationalisations of the common reflection categories.

The chosen setup of the research design allows testing the performance of several types of machine learning under the same conditions because all operate on the same set of data. This allows the comparison of the different types of machine learning algorithm to detect reflection. This also allows comparing the performance of the machine learning algorithms over a variety of datasets important for reflective writing. In the first case, the machine learning algorithms vary, but not the data set; in the second case, the machine learning algorithms are the same, but the datasets vary.

The answers to these two research questions provide empirical evidence for, or against, the automated detection of reflection in texts. The results of this research design are outlined in Chapter 6 'EVALUATION'.

4.7 SUMMARY

This Chapter 4 'METHODOLOGY AND RESEARCH DESIGN' described the methodological considerations made for determining the methods used and the research design for the automated detection of reflection.

The chapter started by outlining the general methodological considerations in Section 4.1 'General methodological considerations'. There, the case was made that, for the aim of this thesis, a quantitative study type that uses machine learning is appropriate. Section 4.2 'Evaluation criteria and metrics' described the criteria for the evaluation of the research question, which are reliability, validity, and objectivity. The criteria and their empirical measurements were described and discussed. The metrics important for the evaluation section of the thesis were determined. Section 4.3 'Unit of analysis' provided guidelines for the proper choice of the analysis unit, and a rationale for the choice of the unit used in this thesis. Sampling techniques were discussed in Section 4.4 'Sampling'. Section 4.5 'Machine learning algorithms' provided the rationale behind the selection of machine learning algorithms in order to

determine good candidates for the automated detection of reflection. Three groups of machine learning models were considered closely: tree-based, rule-based, and several high performing algorithms. For each of these groups, several implementations exist. The concrete choice of implementation was described in Section 4.5.1 'Tree-based models', Section 4.5.2 'Rule-based models', and Section 4.5.3 'High performance models'.

All these considerations were the necessary groundwork that shaped the decisions made for the research design of this thesis. Section 4.6 'Overview of research design' provided an overview of the research design. Section 4.6.1 'Dataset generation process' outlined the process for generating the dataset necessary for machine learning. Finally, Section 4.6.2 'Research design' described the research design and all the decisions that led to it. This section also showed how the research questions are mapped to the research design.

This chapter provided the methodological blueprint for the following two chapters. The implementation of the data generation process is described in Chapter 5 'DATASET GENERATION' and the research design for the evaluation is implemented in Chapter 6 'EVALUATION'.

5

DATASET GENERATION

This chapter describes the concrete implementation of the dataset generation process outlined in Section 4.6.1 'Dataset generation process'. A graphical representation of this process can be found in Figure 2.

The data generation process starts with a text collection and ends in datasets of sentences annotated according to the categories for reflection detection.

In Section 5.1 'Identification of text collection', several text collections were weighted against each other in order to identify the primary source of texts. Thereafter, Section 5.2 'Sampling text collection' outlined the strategy used to sample the text collection to retrieve a suitable subset of texts. These subsets were then divided into units with the approach explained in Section 5.3 'Unitising text collection'. This process resulted in a set of sentences that had to be labelled. These labels supervise the chosen machine learning algorithms (see Section 4.5 'Machine learning algorithms'). The specific approach selected to annotate the collection of sentences is described in Section 5.4 'Overview of annotation task'.

5.1 IDENTIFICATION OF TEXT COLLECTION

The requirements for text collection outlined in Section 4.6.1 'Dataset generation process' were that such collection should be of a reasonable size, and that it should be close to the context of the research on the assessment of written reflection.

Previous work of the author made use of a blog corpus (Ullmann et al., 2012) and data extracted from the forum posts of a virtual learning environment (Ullmann et al., 2013). The blog corpus was dismissed as a choice for this research because the collected blog posts are not related to an academic context. The forum post dataset was also excluded. It contained a wide variety of contributions, which made it difficult to retrieve a suitable collection of texts.

Three corpora of student writings were identified: British Academic Written English (BAWE) (Nesi and Gardner, 2012; Gardner and Nesi, 2013), Uppsala Student English (USE) (Axelsson, 2000), and Michigan Corpus of Upper-Level Student Papers (MICUSP) (Römer and O'Donnell, 2011). They are all good candidate text collections for the aim of this thesis because of their size and relatedness to the academic context in which most of the research on the analysis of written reflection was conducted. The author decided that the BAWE corpus should be one of the main sources of the text collection.

BAWE[1] is a text corpus of 2,761 assessed undergraduate and postgraduate university student text files from 35 disciplines in the four general areas of Social Science, Arts and Humanity, Life Science, and Physical Science (Nesi and Gardner, 2012; Gardner and Nesi, 2013). The corpus is freely available, and its license permits its usage for research[2].

According to the creators of the BAWE corpus, one of the aims was to assemble a balanced dataset with a broad disciplinary scope to capture student writings (Gardner and Nesi, 2013, p. 32).

1 The BAWE corpus was developed at the Universities of Warwick, Reading, and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly from the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from ESRC (RES-000-23-0800).

2 The corpus is available from the University of Oxford text archive: http://ota.ahds.ac.uk/headers/2539.xml.

It is notable that the texts of the corpus are annotated with a rich set of metadata, which allows the inspection of various aspects of the dataset, as well as cross-comparisons. Examples of these metadata are:

- Student information, e.g., unique identifier, gender, year of birth, first language, and education.
- Study level (first year, second year, third year undergraduate, postgraduate).
- Grade: distinction (grade 70% or higher) and merit (60% to 70%).
- Discipline (with sub-disciplines and modules).
- Genre (e.g., essay, critique, narrative recount, etc.).

The next step of the data generation process is to sample the text collection (see Section 4.6.1 'Dataset generation process').

5.2 SAMPLING TEXT COLLECTION

As indicated in Section 4.6.1 'Dataset generation process', manual inspection of the BAWE corpus shows that it contains reflective texts, but also that most of the text is classic student academic writing, written in a factual style without any reflection. In order to find a suitable sub-sample of the entire corpus, relevance sampling mostly based on several objective criteria is chosen. These criteria are used to triangulate a list of texts relevant for this thesis.

With regard to the generalisability of the results, a random selection of texts would have been the ideal sampling technique in order to retrieve a sample of texts (see Section 4.4 'Sampling'). In a previous work of the author (Ullmann et al., 2013), a randomised sub-sample was obtained from a large text collection of forum posts from a virtual learning environment. In this case, the sampling technique was suitable because the preliminary analysis of the text collection indicated that the sentences, which were of a personal nature, were sufficiently equally distributed over the entire text collection, which made it suitable for random selection.

However, such is not the case for the BAWE corpus. Preliminary analysis of this corpus indicates that it contains many factual student writings as assignments are mostly on discipline-related topics than reflections of personal experiences. Thus, a random selection of texts would have led to many texts unrelated to reflection. As described in Section 4.6.1 'Dataset generation process', the chosen sampling technique is relevance sampling. This is a technique based on explicit relevance criteria to retrieve texts that are relevant for answering the research question. For this research, the texts of relevance are reflective writings. The criteria that led to the inclusion of certain texts into the final collection of texts are described here.

Nesi and Edwardes, one of the creators of the BAWE corpus, indicated that the corpus contains writings of several reflective writing tasks (Nesi and Edwardes, 2007). The paper of Nesi and Edwardes (2007) highlighted eight text snippets as examples of such tasks. The inspection of the metadata of the corpus showed that some of the texts are marked as texts that contain reflective recounts[3]. For example, a text can be a compound of creative writing and reflection, or an essay and reflection (see also Heuboeck et al. (2007, p. 21) for a detailed overview of the corpus metadata). Texts that contain a reflection task according to the corpus metadata form the first set of reflective writing candidates for the sampled text collection. However, the list was relatively small and other heuristics were employed to extend the list of candidate writings.

An indicator of reflective writing can be the title of the text, for example, if it contains references to reflection or experience. Regular expressions are used to find titles that refer to reflection and experience. From the 2,761 titles of the BAWE corpus, 43 titles contain 'reflect*' (i.e., the root word 'reflect' and other forms of the word, such as reflection, reflected, reflecting, reflective, and more). Among them are titles

3 This is only visible in the XML version of the corpus, more specifically, in the 'note' section of such version. A regular expression was used to search all XML files with regard to the text 'reflect*' in the 'note' section of the XML. An example of this annotation is: 'evaluated as candidate compound assignment. Assigned to A4: text + major reflection [task]: compound'.

that indicate a reflective writing, such as 'Individual Reflective Assignment', 'Individual Reflective Piece', 'Reflect on a challenging experience in relation to the practice...', or 'A reflective account on the process of teamwork'. However, there are also titles that are not indicative of reflective writing, for example 'How are the functions of Roman towns reflected in the Archaeological record?', 'Examine diplomatic methods used by Russian Fed in the conduct of its FP. In what ways, if at all, does it reflect dom interests?', or 'With particular reference to the work of Jacques-Louis David, how were the principles of the French Revolution anticipated, reflected and promoted in French painting, 1784-1794?'.

A total of 24 titles contain references to experience (e.g., experience, experiences, experiencing, etc.). Among them are candidates for reflective writing, for example, the titles 'Describe and analyse an ethical dilemma you have encountered in your own professional experience', 'Professional management experience - Year-out work experience', or 'Motivation, Input and output in my Persian Learning Experience'. However, there are also titles not related to reflective writing, such as 'In what sense does our experience of the sublime demonstrate the supremacy of our rational nature, according to Kant?', 'How are poor household's experiences of food shaped by their social, economic, or personal experience?', or 'Nicaraguan experience of agricultural and rural development strategy in the period from 1979 to 1990'.

When considering indicators at the text level, we can argue that an indicator for reflective writing might be text written from a personal perspective. Indicator words for this can be 'I', 'me', or 'myself'. The following two figures show relative frequencies for all texts: Figure 5 shows the relative frequencies of the character 'I', and Figure 6 shows the relative frequencies of sentences that contain a first-person singular pronoun.

The first heuristic uses a regular expression to search for occurrences of the character 'I' in the text. The 'I'-count divided by the character count of the text is equal to the

relative frequency. The 'I' character mostly designates the first person pronoun, but there are also exceptions, for example, in abbreviations such as 'A. I. Thompson'.

The second heuristic uses a rule to infer whether a sentence contains at least one first person singular pronoun (such as 'I', 'me', 'myself'). The following listing describes the rule in natural language. This rule was developed as part of an extension to the author's reflection detection architecture (Ullmann, 2011; Ullmann et al., 2012).

Listing 1: Personal sentence rule

```
Rule 'personal_sentence'

FOR ALL sentences in document
IF
        Exists a personal pronoun OR a possessive pronoun
        AND
        this pronoun belongs to the dictionary of self-referential pronouns
THEN
        Mark sentence as relevant
```

Similar to the first heuristic, the relative frequency is calculated by the count of relevant sentences divided by all sentences in a text.

We can see from both figures that more than half of all BAWE corpus texts do not contain any of the two indicators for personal texts. Further, the number of texts with proportionally more signals that indicate personal text decreases rapidly. However, this also shows that personal texts do exist.

The top text in both figures starts with: 'I have now completed a year's work experience at Cowley Manor hotel, and feel I have developed as a person. Moving into the world of full time work after my first year at university was quite a shock. The first few weeks were busy with me settling into a new job and new town. I soon discovered that I enjoyed the set routine that comes with a full time job and knowing that time outside work was my own. I have always thought of myself as an independent person, but feel that my time in Cheltenham made me more independent. I looked after myself, and in work, I made a point of being eager to learn, and not be "spoon fed." I feel that I swiftly settled into work. This is reflected in my first appraisal, when Jake mentioned that I had adapted well to the areas I had

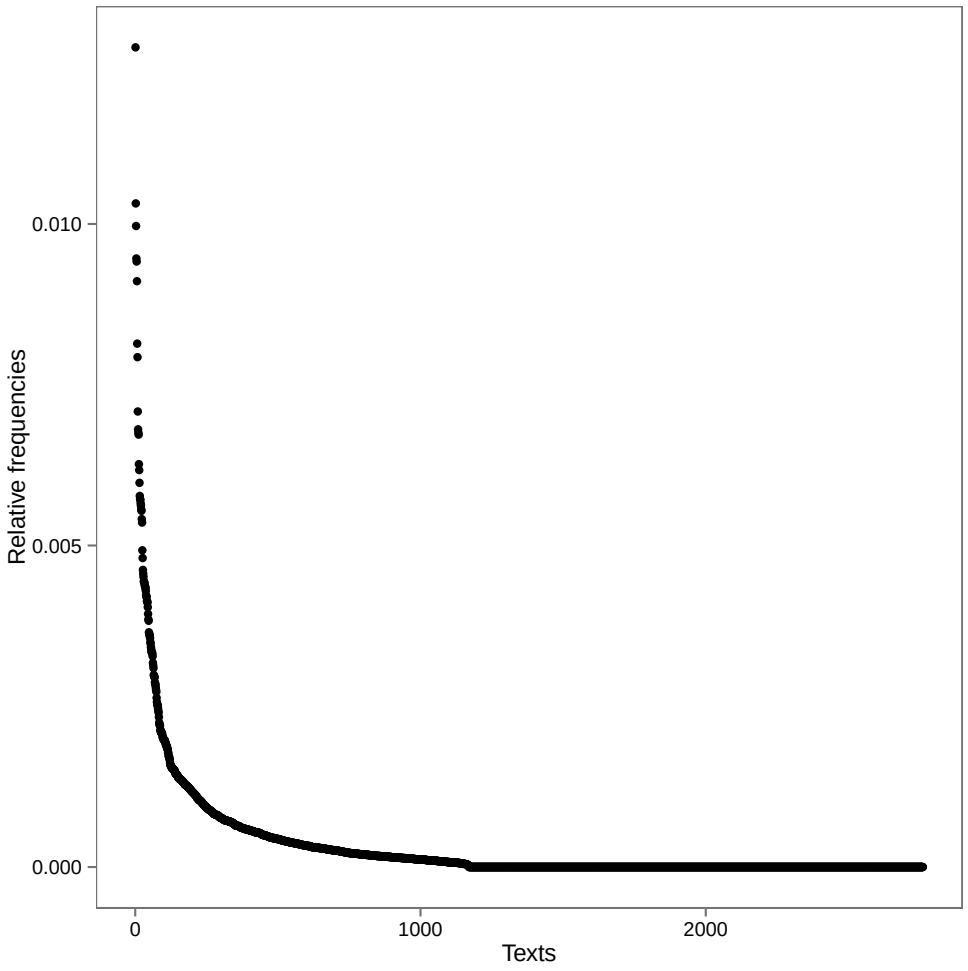

Figure 5: Relative frequencies of 'I' for all BAWE corpus texts

worked in. When reflecting upon my first appraisal, I felt that my efforts to integrate myself and settle in had been recognised (...)'.

Let us compare this top text with the text on position 100 with regard to the relative frequency of 'I'. The text starts with: '"Pan Recipe" is a Caribbean poem written as an extended metaphor, which uses the vehicle of the steel pan to convey a sense of new life born out of the past, and is suggestive of the oppression of black slaves. When I began writing "How we have walked, How we have journeyed", I aimed to re-centre Agard's text and focus more primarily on the Caribbean music itself, as a celebration of freedom, but still express ideas about the oppression of black slaves through metaphor. (...)'.

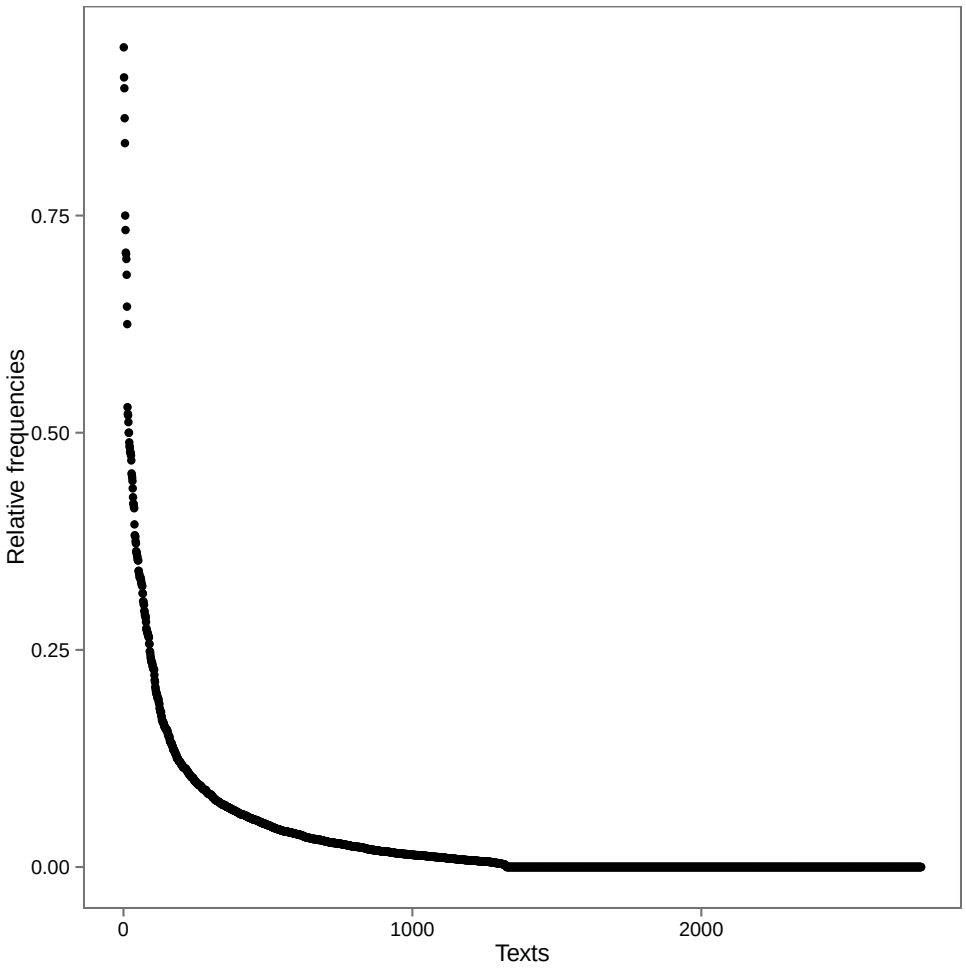

Figure 6: Relative frequencies of 'personal' sentences for all BAWE corpus texts

The following is an example of text that contains no reference to the character 'I': 'The history of English law is long standing and well established. As stated in Keeton's (1984:7) book, "the doctrine of precedent inherently brings legal history to bear upon current judicial decisions." Due to the distinctive nature of precedents, the specificity of cases had been coloured by its uniqueness and independence. (...)'.

In the top text, nearly every sentence has a reference to the first-person singular pronoun, and the character 'I' occurs frequently. It is an account written throughout from the perspective of the writer. The other two examples are more descriptive, less pronounced with regard to the recount of a personal experience. The manual inspection of the top texts derived with the first and second method indicate that after 100 texts, mostly descriptive, factual text can be found.

Hitherto, we have discussed four indicators of reflective writings: the title information, relative frequency of first-person singular pronouns in texts, texts mentioned by Nesi and Edwardes (2007), and references to reflection in the markup of the BAWE corpus. All these indicators were taken to generate a list of candidate texts that might contain reflection. The list contains all texts with 'reflection' and 'experience' in the title, the top 100 texts identified with the first heuristic, the top 100 texts identified with the second heuristic, the texts mentioned by Nesi and Edwardes (2007), and the texts that contain markup that indicates a reflection. The process described hitherto resulted in a set of 141 unique texts (out of the original 2,761 texts).

These texts were then inspected manually, and such inspection revealed, first, that not all texts were good candidates because they were largely written in a factual style (see the third example above as an illustration of such style). For the most part, these texts originated from the automated sampling method. Second, some texts contain only a small proportion of an experience description. Often, only one of several tasks in a writing assignment asks for personal reflections (compound assignments). Third, while these designated sections aim at reflective writing, this does not necessarily mean that the students responded with personal reflections. These insights led to the decision of reducing the amount of texts and prune those texts that contain long passages of factual writings. Therefore, texts were maintained, deleted, or pruned. Appendix C contains a list of all texts and the decision made for each text. Those that were deleted were largely descriptive. For example, one person wrote a step-by-step description of the development of a computer program. Another text was a linguistic study of the word 'I'. Another text mostly contained a fictional story narrated in the first person. Some texts only contained small sections on personal experience; these were pruned. In total, this process led to a text sample of 67 texts.

In addition to these 67 texts, ten other texts were included that were selected from the literature on reflective writing. These texts are distinguished reflective writings,

and they were added in order to have, in addition to the texts of the BAWE corpus, texts directly stemming from the literature on reflective writing. Not many such texts are available publicly, and thus only ten are included. These texts are 'The park account 4', 'the Presentation account 3', 'A dance lesson account 3', 'Placement on business management programme account Barry', 'Reflection on study habits account Kerry', 'A GPs story account 4', 'the worrying tutorial account 3', 'the surprise at home account 3' from the resource sections of Moon (2006) and Moon (2004), the logbook excerpt from Korthagen and Vasalos (2005), and the reflective account described in the supplementary material of the paper for Wald et al. (2012). These are reflective texts that were thoroughly analysed with regard to their reflectiveness, and such analysis can be found in the cited literature.

The described relevance sampling approach identified 77 texts that are seen as relevant texts for the aim of this thesis. The final collection of texts is based on the triangulation of four objective criteria that identify the texts automatically, the perspective of other researchers working with the corpus, and the author's experience to select the final text collection. The list of candidate texts can be replicated by other researchers with the information provided above. All decisions with regard to the selection of the final text collection based on the candidate lists are made explicit in Appendix C 'SAMPLED TEXT COLLECTION OF THE BAWE CORPUS'.

These texts are mostly those that contain a personal narrative. The chosen criteria for the relevance sampling are aimed at raising the likelihood that the selected texts contain reflection. This is especially important considering that only few texts from this large corpus are relevant to this research. A random sampling strategy would have significantly increased the annotation effort in order to find evidence of reflection. Here, a relevance sample is more suitable.

The 67 texts selected from the BAWE corpus were written by 47 students: 12 texts were written by first year students, 21 by second year students, 23 by third year

students, and four by postgraduate students. For one text, the student level was unknown[4]. Most texts were graded with merits (N = 40), followed by distinction (N = 24). The grade is unknown for three texts. Most texts are from the disciplines of Health, followed by Business, and Engineering (see Table 5). It is notable that the discipline Health contained many more texts than the other disciplines. Mann et al. (2007) noted that reflective practice is an increasingly important part of the medical education curriculum.

Discipline	Frequency
Health	20
Business	9
Engineering	9
Hospitality, Leisure, and Tourism Management	6
Linguistics	6
Sociology	3
Computer Science	2
Cybernetics and Electronic Engineering	2
English	2
Other	2
Philosophy	2
Archaeology	1
Law	1
Medicine	1
Politics	1

Table 5: Distribution of disciplines

Dividing the texts into analysis units is the next step of the data generation process, as outlined in Section 4.6.1 'Dataset generation process'.

5.3 UNITISING TEXT COLLECTION

The choice as unit of analysis felt on sentences (see the rational for this in Section 4.6.1 'Dataset generation process').

4 This information is obtained from the metadata spreadsheet that accompanies the BAWE corpus.

Text can be divided automatically into sentences. There are many ways of doing so. A simple strategy is to divide texts at full stops or question and exclamation marks with regular expressions. However, this is prone to errors, for example, the sentence 'Mr. Thomas D. Ullmann presents reflection detection.' might be divided into the parts 'Mr', 'Thomas D', 'Ullmann presents reflection detection'.

A better method for extracting sentences from texts is to use a machine learning model trained to detect sentences. A well-trained model can detect the sample sentence above in its entirety. The model used[5] has high accuracy for detecting sentences. A program was written that splits the texts into sentences.

The result of this process to generate units from the 77 texts is a total of 5,131 analysis units. Such units are yet to be labelled. The next section describes the annotation task.

5.4 OVERVIEW OF ANNOTATION TASK

The annotation task step of the data generation process, as outlined in Section 4.6.1 'Dataset generation process', consists of pilot and annotation phases. Pilots were arranged in order to determine the final task design of the annotation task. As outlined in the data generation process, an approach had to be found to scale to the size of the text collection. The chosen approach is to 'crowdsource' the annotation task. Before outlining the pilots that led to the final task design, the next section outlines relevant background information on large-scale crowdsourced text annotation tasks.

Section 5.5 'Background on crowdsourced text annotation provides an overview of the research that used crowdsourcing to annotate textual data. Most importantly was the question of whether a crowdsourced approach can produce high quality annotations of textual data. The research outlined in Section 5.5.1 'Research on

5 Based on the Java library Apache OpenNLP (https://opennlp.apache.org).

crowdsourced annotation quality' suggested that crowdsourcing is a good candidate for large-scale text annotations. This also highlighted the importance of a carefully designed task design (see Section 5.5.2 'Research on crowdsourcing task design'), and further, that the best results can be achieved if multiple ratings are aggregated (see Section 5.5.3 'Aggregating crowdsourcing results'). Because the research on crowdsourcing is only on related problems, and not on the problem of annotating text with regard to reflection, a series of pilots were conducted in order to evaluate the potential of crowdsourcing for annotation with regard to the common categories of reflection. The pilot results are summarised in Section 5.6 'Summary of pilots'. The experiences made during the pilots significantly influenced the task design used to annotate the text collection. The annotation task is described in Section 5.7 'Annotation task'. The reliability of the raters to annotate sentences with regard to indicators of the common categories of reflection is outlined in Section 5.7.4 'Reliability'. In addition, the validity of the indicators is confirmed empirically in Section 5.7.5 'Validity'. The outcome of the data generation process is, for each common reflection category, a dataset of highly agreed sentences. The applied quality standard and the descriptive dataset statistics can be found in Section 5.7.6 'Quality standard and datasets statistics'.

5.5 BACKGROUND ON CROWDSOURCED TEXT ANNOTATION

As outlined in Section 3.1 'Manual methods to detect reflection', one of the fundamental tasks in assessing reflective writings is to analyse texts using a coding schema that assigns several reflection categories to text-segments. This task is usually performed by a small group of trained coders (usually two to three coders) that analyse a small number of texts.

Using a crowdsourcing approach is a promising method for scaling the task to thousands of coders who label text. Crowdsourced annotation tasks are seen by many researchers as a suitable way for gathering large annotated datasets for further research.

There are many forms of crowdsourcing and motivations to participate in crowdsourcing (Wang et al., 2013; Yuen et al., 2009; Quinn and Bederson, 2009, 2010). Some prominent examples are paid crowdsourcing platforms, games with a purpose, and the Wikipedia platform.

For this research, the focus is on crowd-worker platforms[6]. These are crowdsourcing platforms that help distribute tasks to a large amount of workers paid for their services. Usually, these tasks are rather simple and can be completed in a short time.

There is already a growing body of research in the area of crowdsourcing and its challenges (Kittur et al., 2012). This thesis focusses on a specific domain of crowdsourcing, the annotation of texts. The manual rating of texts related to reflection categories can be described as a task that involves a subjective rating of text segments. The following sections present research that describe the results from crowdsourcing experiments on text ratings according to subjective criteria. This is in contrast to tasks where the results can be objectively verified. Examples for the latter case are counting pictures on web pages, or determining whether the pictures show a person.

5.5.1 *Research on crowdsourced annotation quality*

Kittur et al. (2012) stated that quality control is one of the main research issues of crowdsourcing. There are mixed reports on the quality of crowdsourcing that vary from research that attested crowdsourcing to have the same quality as expert ratings if

6 Examples are CrowdFlower http://www.crowdflower.com/ or Amazon's Mechanical Turk https://www.mturk.com/mturk/welcome.

the results are aggregated (Li et al., 2013, p. 1), to articles that report a high amount of low quality work (Bernstein et al., 2010, p. 316).

This section reports on research that investigates the quality of crowdsourced work on tasks, where the workers had to rate text according to several characteristics. This type of task is related to the task of annotating text segments according to reflection elements. All the work described below compares the work of the crowd with experts. The aim is to gain an understanding on the quality of crowdsourced data gathering compared with domain experts.

Snow et al. (2008)[7] presented research in the area of natural language understanding. Snow et al. (2008) ran several tasks on Mechanical Turk, and compared the results of the crowdsourced workers with expert ratings of the same data.

The first task was to rate short headlines with respect to the dimensions of anger, disgust, fear, joy, sadness, surprise, and valence. Snow et al. (2008) compared the quality of expert and non-expert labels in two different ways. First, Snow et al. (2008) compared individual expert coders with individual crowdsourced coders. Snow et al. (2008) found that the agreement between experts is higher than the agreement between non-experts and experts. Subsequently, the authors aggregated the results of the non-experts in order to assess how many non-expert coders are necessary to achieve the results of the expert. As an aggregation strategy, they used the average of all raters. Snow et al. (2008, p. 257) stated that on average, four non-experts are necessary to reach equal or better agreement than the expert annotators.

The next task was a similarity task where pairs of words were evaluated according to their relatedness. The aggregated results of ten workers – again, using the average of the ratings – reached a Pearson's r correlation of 0.952, which is close to the correlation scores of previous lab research (Snow et al., 2008, p. 257 f.).

7 Data available at https://sites.google.com/site/nlpannotations/

The third task was on recognising textual entailment. The workers had to evaluate two sentences, and determine whether the second sentence could be inferred from the first. Snow et al. (2008, p. 258) aggregated the results of ten workers with the simple majority vote ('tiebreakers' were used randomly for equal ratings). For 100 sentence pairs, they found an agreement of 89.7%, which is close to the expert agreement reported in the literature (Snow et al., 2008, p. 258).

The fourth task consisted in determining whether an event occurred before or after another. The aggregation of ten annotators with simple majority vote reached an agreement of 0.94 with the gold standard annotations (Snow et al., 2008, p. 258f.).

The last task on word sense disambiguation also found high agreement (0.99) between the aggregated non-expert and the gold standard (Snow et al., 2008, p. 259).

The results from Snow et al. (2008) suggest that for many different tasks, the aggregated ratings of many non-experts are comparable to expert ratings.

Kittur et al. (2008, p. 454-456) described two experiments with Mechanical Turk[8] in the context of quality ratings of Wikipedia articles. The task for the workers consisted in reading preselected Wikipedia[9] articles, and then rating them according to their quality. The worker ratings were compared with the ratings of Wikipedia administrators (experienced Wikipedia users). The authors stated that the questions for the ratings were oriented on the criteria for featured articles[10]. The authors asked the workers to rate how well the article was written, how accurate, whether it was neutral and well structured, and the overall quality of the article on a seven-point Likert scale. In addition to rating the articles, the workers were asked to complete a text-field with suggestions for improvements for the article. A total of 15 workers rated each article. A marginally significant correlation coefficient of $r = 0.5$ ($p = 0.07$) between the worker ratings and the Wikipedia experts was reported. The authors

8 https://www.mturk.com
9 http://en.wikipedia.org
10 http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria

found that, from 210 free-text responses, 102 were outside the task, and 64 ratings were completed in an extremely short time (under 1 minute). The authors outlined that a small group of workers was responsible for a large amount of invalid ratings.

The second experiment used a set of test questions (questions with known answers) before the workers actually saw the Wikipedia article. The questions were on how many references, images, and sections the article had. With this test only seven out of 277 responses were outside the task. The rating question was on the overall quality of the article measured on a seven-point Likert scale. The significant correlation between the workers and Wikipedia administrators was $r = 0.66$ ($p = 0.01$).

The Kittur et al. (2008) task seems more difficult than the task reported by Snow et al. (2008) because the former asked how well an article was written, which can be seen as a rather subjective task. In addition, the workers had to assess an entire text, compared with small text-segments. Although working with full text, the authors found a relatively strong correlation between the workers and Wikipedia administrators after introducing the test questions.

Alonso and Mizzaro (2012) used crowdsourcing for relevance assessment of TREC data. TREC is the text retrieval conference that hosts competitions related to information retrieval. In many cases, the competitions are based on a gold standard dataset generated manually by human coders. Data generation is expensive, and therefore, other ways of annotating data are explored, including crowdsourcing. In their paper, the authors examined the degree to which TREC assessors can be replaced by crowdsourcing. They concluded that crowdsourcing is a reliable alternative for relevance assessment (Alonso and Mizzaro, 2012, p. 1053). The task for the worker consisted on rating documents according to their relevance to a topic description. The task used a binary rating system (relevant vs. not relevant). Five workers rated each document-topic pair. The task contained test questions. The crowdsourcing results were then compared with expert assessor ratings.

For the task, the authors reported individual and aggregated agreement metrics. Individual agreement is the agreement between each worker and TREC assessor. In 68% of the cases, the workers agreed with the TREC assessor, which the authors determined as fair, but not great, agreement (Alonso and Mizzaro, 2012, p. 1057).

The authors then aggregated the worker results using majority vote (three out of five) in order to compare with the assessor. This led to 77% of cases agreeing with the assessor.

The authors also compared the agreement values with the values reported in other research on relevant assessment. They stated that the group agreement of 77% of the crowd workers is comparable to the 78% agreement of expert coders reported in other literature on a binary task.

In their conclusion, the authors stated that their results need further confirmation, but 'they support the idea of using a group of workers to gather redundant judgments, and they are a good indication that MTurk can be a reliable, quick, and cheap alternative for relevance assessment in TREC-like initiatives' (Alonso and Mizzaro, 2012, p. 1064-5).

The research of Alonso and Mizzaro again emphasised the importance of aggregating the ratings of many raters.

Schnoebelen and Kuperman (2010) compared the quality of crowd workers with the quality of traditional lab experiments in the area of linguistic research. They concluded that 'Mechanical Turk is a reliable source of data for complex linguistic tasks (...)' (Schnoebelen and Kuperman, 2010, p. 441).

Their first experiment was the Cloze sentence completion task. Workers were given part of a sentence, and asked to write the word they believed would follow next. Each task consisted of 12 stimuli, and was completed by 50 workers. For this task, the authors reported a strong correlation (Spearman's rho = 0.823) between lab participants and crowd workers.

The second set of experiments compared the performance of workers with students. The task was to rate the similarity between verbs and verbs in phrases (phrasal verbs) on a seven-point scale (Schnoebelen and Kuperman, 2010, p. 451 ff.). A total of 18 and 27 ratings per item were collected for both student experiments, and approximately 29, 25, and 19 ratings per item for the three crowdsourcing experiments. The correlation between the workers and students was high in the context experiment (rho = 0.9) (Schnoebelen and Kuperman, 2010, p. 453) with a high Cohen's κ of 0.823 (Schnoebelen and Kuperman, 2010, p. 457). The correlation between the three crowdsourcing experiments was high (rho > 0.7) (Schnoebelen and Kuperman, 2010, p. 453), and the correlation between the student experiments was relatively low (rho = 0.44) (Schnoebelen and Kuperman, 2010, p. 453). The authors stated that the worker data were more consistent than the student data (Schnoebelen and Kuperman, 2010, p. 457).

As with the other studies, Schnoebelen and Kuperman (2010) reported a high correlation between crowd coders and traditional coders. However, they did note that lab participants performed better on the word completion task than crowd workers. In the second set of experiments, which used a rating scale, the authors reported high correlations between workers and student participants, and the results of the workers were more consistent than those of the students.

Benoit et al. (2012) reported results from a crowdsourcing study in the area of political texts. They compared results of expert coders with the results of workers, and concluded that if the task is designed carefully, crowdsourcing can replicate the results of expert coders (Benoit et al., 2012, p. 24). The task was on coding sentences from the political manifestos of three British parties with regard to economic policy (left vs. right) and social policy (liberal vs. conservative) on five-point scales (Benoit et al., 2012, p. 7 f.). The goal of their research was to detect party shifts over points in time. The shifts in question were already described in other studies, and are thought

to be replicable with expert and crowd coders. The results were that six expert coders (three were authors of the study) could replicate the results of previous studies, as could the crowd workers. The authors stated that there is a strong and significant relationship between the workers and experts (Benoit et al., 2012, p. 23) and that '(...) non-expert coders scattered around the world are generating essentially the same estimates of party policy positions as 'expert' political science professors and graduate students' (Benoit et al., 2012, p. 22). Although the data were noisy at the individual level, the aggregated worker results were useful.

The paper from Benoit et al. (2012) stated a similar conclusion to the research outlined above, where crowdsourced work can be of good quality if the worker results are aggregated.

Overall, this synthesis of the literature on the research of paid crowdsourcing suggests that the aggregated results of tasks are positively related to the performance of experts, and can come close to expert ratings. Although the research outlines that crowdsourced work contains noise that negatively impacts the quality of the work, there are sufficient signals to achieve results that are comparable to expert ratings.

Although the results above are encouraging, coder expertise is not unimportant. Results from the area of relevance assessment for information retrieval indicate that different levels of expertise indeed influence measures in a small, but consistent, way (Bailey et al., 2008).

The next section highlights important design decisions found in the research of crowdsourcing text annotations.

5.5.2 *Research on crowdsourcing task design*

The careful design of crowdsourcing tasks seems to be the key for generating data of high quality.

Several researchers mentioned embedding test questions into the task. Test questions can be seen as a way of training workers on the task, and as a test to check whether workers understood the task sufficiently well. For example, Benoit et al. (2012, p. 12) reported that their task contained several test questions (gold questions with known answers) to ensure a minimum qualification level of the workers for this task. Each target sentence was rated by at least five workers. The observed patterns were more visual if the dataset was filtered for trusted raters. Trust values were provided by the crowdsourcing platform of their choice. The authors defined a threshold value of 0.70. Workers with higher trust values than the threshold were defined as trusted workers. These test questions are seen as the key to reliable data (Benoit et al., 2012, p. 12). In addition to the test questions, the authors deployed 'screener' sentences that served as indicator of the workers paying attention to the task (Benoit et al., 2012, p. 13). Screener sentences prompted workers to generate a specific response in the rating scales. If this response was not made, this was an indication that the worker did not pay attention to the task. In addition, Kittur et al. (2008) indicated the importance of the test questions, and that the task should be designed in such a way that arbitrary answering is more difficult than answering in good faith.

Schnoebelen and Kuperman (2010, p. 462 ff.) provided several recommendations for designing a crowdsourcing task and cleaning data. The following is summary of some of the recommendations that are relevant for this research: start testing the task with small pilots and keep the task short. Define exclusion criteria for filtering the results. Avoid 'cherry-picking'. For tasks in English, choose workers from English speaking countries. If an answer has to be provided in a specific format (for example, write only one word), the responses that do not fulfil this format can help exclude workers. The time spent on a task is an indicator to remove workers. Schnoebelen and Kuperman removed those workers with more than two standard deviations from the mean. On average, the authors cleaned 25% of the workers (37% of the data points).

5.5.3 *Aggregating crowdsourcing results*

Wang et al. (2013, p. 10) saw crowdsourcing as an alternative to achieving high quality annotations because many redundant annotations would help filter noise. A common way of aggregating data is by majority voting, which is not perceived as being perfect, but can generate satisfying results (Li et al., 2013, p. 2).

Hovy et al. (2013, p. 1120) argued that majority voting is deficient when many coders always choose the same category dimension (for example, in the case where the workers always rate all categories as yes). The authors used an item-response model to find latent true labels of items from messy data. Their model outperforms several other models (Hovy et al., 2013, p. 1124) with regard to accuracy, including majority vote, if applied to the natural dataset of Snow et al. (2008). However, these differences are small. The model achieves better accuracy with an increasing number of annotators (Hovy et al., 2013, p. 1126). On a synthetic dataset that simulates two gamer strategies (a random rater and a rater who always rates the same category) and different gamer percentages (from 50% to 100% of gamers), their model performed better than applying majority vote as aggregation strategy (in the range of 50% to 70% of gamers).

Gao and Zhou (2013, p. 8 f.) remarked that, when comparing majority vote with the Dawid-Skene estimator for aggregating rater results, the Dawid-Skene estimator is better than majority vote in the case of having mostly results from gamers.

5.5.4 *Summary*

The research on crowdsourcing annotation tasks is a relatively new development that bears the potential of annotating large amounts of data in a reasonable time.

From the research on the quality of crowdsourcing in related areas outlined above, it seems that crowdsourcing is a viable method that, when carefully designed and if the results are cleaned and aggregated, may be close to expert evaluation, or to the quality of traditional lab experiments.

Furthermore, the description of crowdsourcing tasks contains all the information required to replicate the task. Such description contains coding instructions and the test questions required to qualify the task annotators. This strengthens the reproducibility of the data generation process because all the information is made explicit (see the evaluation criteria of objectivity in Section 4.2 'Evaluation criteria and metrics'). The manual content analysis of reflective writing often relies on coder training. Explicit description of coder training and the process for developing the mutual understanding of how to code text is often not available. Poldner et al. (2012, p. 31) found in their review that only one out of 17 articles described the coding protocol.

5.6 SUMMARY OF PILOTS

Several pilots were conducted before the actual annotation task. As described in Section 5.5.2 'Research on crowdsourcing task design', the task design is crucial in order to achieve high quality annotations. This section summarises several observations made with experiments that explored the potential of using paid crowdsourced workers to annotate texts and text-segments according to important elements of reflection. These observations influenced the final task design used to create datasets for the evaluation part of this thesis.

The task design building blocks are discussed here. We start with the observation that, although most crowdsourced coders work diligently, a certain amount of crowdsourced coders do not consider the task important. Section 5.6.1 'Gaming behaviour' raises awareness on this issue because neglectful gamers is detrimental to

the quality of the annotations. A novel approach to significantly reduce the amount of gamers is presented in Section 5.6.2 'Custom validators'. Section 5.6.3 'Amount of ratings' discussed the amount of coders necessary to rate one unit. Section 5.6.4 'Multiple-choice vs. rating scale' discussed issues regarding the scale most suitable to the task. A summary of all recommendations closes the discussion on the decisions made to design the task for the data generation annotation.

5.6.1 *Gaming behaviour*

Early experiments of the author that used crowdsourcing to annotate texts emphasised that a certain amount of workers (people who perform the tasks on crowdsourcing platforms) attempt to game the task in order to be paid with minimal effort. A worker is paid immediately after completing a task. This means that one could click through a task as quickly as possible to fulfil the task, without reading the task description, the task, or the item of the task. In this thesis, we refer to these workers as 'gamers' because they attempt to 'game' (manipulate) the task. This follows the term used by Kittur et al. (2008). Other people refer to such workers as spammers, for example, (Li et al., 2013, p. 2), (Hovy et al., 2013, p. 1120), and Gao and Zhou (2013, p. 1). Detecting gaming behaviour is important because gaming negatively affects the reliability and validity of the annotation task. If gaming can be detected, these workers can be filtered, which in turn improves the quality of the task results. However, gaming is a complex subject. Therefore, several thinking points are outlined and discussed later in this paper.

The time required by a worker to complete a task can be indicative of gaming. A task requires a minimum amount of time to be processed by a worker. The worker has to read the instructions and the text of the task, and then requires time to answer the task questions. Therefore, if the worker completes a task in an extremely short time,

shorter than possible to read the text, this information can be used as an indicator of gaming behaviour.

The following example illustrates this type of gaming behaviour. The task consists of several task pages (for a concrete example of a task set-up, see Section 5.7.1 'Task setup'). After the worker completes and submits the page, he or she receives payment. A task page consists of five sections. Each section contains blog text and several rating questions on the text. There is no limit to the number of pages on which a coder can work. Each item receives at least three ratings. The task does not use any test questions and no free-text fields.

The amount of sections within a page can be specified in the task set-up of the crowdsourcing platform. In general, this is a fixed number of sections. Sometimes, however, the crowdsourcing platform varies the amount of sections in order to gather any outstanding ratings.

To inspect the time a worker dedicated on a task, we can plot the time per page for each worker. The following Figure 7 shows these data ordered from workers who spent the least amount of time on the tasks (for better visibility, 2/3 of the data are shown). For each worker the data are ordered from the first page they visited to the last page. We can see that the coders on the left side of the graph dedicated only seconds to a page, which can indicate that these coders did not read the text thoroughly and rated the items randomly.

The figure reads as follows. The y-axis depicts the time difference between the start time and submission time of each page. The x-axis groups unique raters and pages. For example, r001-p01 represents a rater with the unique ID 1 completing items on page 1. r001-p02 is the same rater for page 2. r002-p01 signifies rater 2 and page 1. For example, rater 1 dedicated most of the time to page 1, and then completed all other pages much more rapidly.

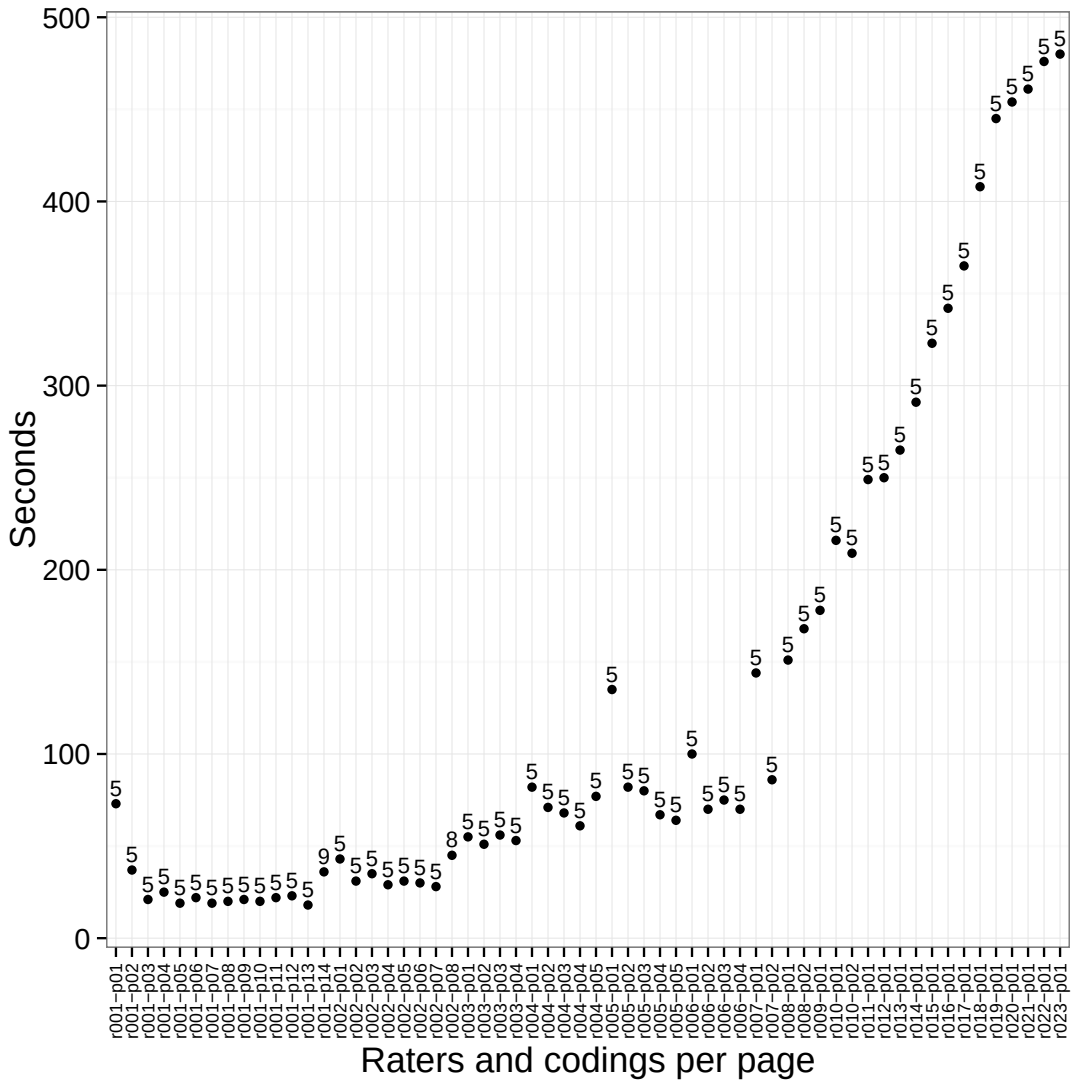

Figure 7: Time dedicated to crowdsourcing task

The numbers above each dot in Figure 7 represent the number of sections a rater submitted at the same time. Mostly, they are multiples of five that reflect the page set-up for the five sections of a page. This number varies sometimes depending on the number of sections allocated by the crowdsourcing platform.

We can see that there are some raters who completed the pages in a very short time. This is an indicator that these workers were gaming the task because it is unlikely that the workers studied and carefully rated five blog posts in such a short time.

Next, we describe how the removal of these gamers influences the overall quality of the task in terms of the per cent agreement of raters. As outlined above, these values

are derived from a task that does not contain test questions to qualify the coders, nor any other measures to improve coding quality.

Before filtering workers based on their time dedicated to the task, we determined that the item 'the text is reflective' has a percentage agreement of 61%. The percentage agreement for the item 'the text contains a description of what was happening' is 62%, and the percentage agreement of the item 'the text shows evidence of a personal experience' is 63%.

When all those coders who dedicated less than 150 seconds to the task are removed (see the gap within the time band in Figure 7 between 100 to 150 seconds), the agreement percentages are higher. Two cases are considered. For the case where all items are removed, which were rated less than three times, the value for the reflective item is 69%; for the description of what is happening, it is 81%; and for the personal experience item, it is 75%.

For the case where all items are maintained (which means that some items received only two ratings, or even one), the value for the reflective item is 67%; for the description of what is happening, it is 71%; and for the personal experience item, it is 68%. Again, the agreement values are higher if those people who dedicated an extremely short time to the task are excluded.

In the following paragraphs, we propose another method for detecting gamers that is used for the task design of the corpus of reflective and non-reflective sentences.

Whereas short time dedicated to a task might be indicative of gaming behaviour, one has to consider that there are also workers who are extremely specialised in the tasks and can rapidly perform the tasks well. Such quick workers, while not as fast as the gamers described above, might be sufficiently quick as to fall within the threshold of two standard deviations from the mean, as described in Schnoebelen and Kuperman (2010, p. 464). This would exclude perfectly good workers, if only task duration is considered for gaming detection.

Such high performance contributors should not be dismissed from the results set, or be banned from the task (banning is one mechanism to exclude workers from following tasks). Another indicator for gaming behaviour can be the free-text answer field. For example, workers are asked not only to provide a rating for an item, but also a short explanation as to why they rated the item in such way. Therefore, the free-text answers can provide qualitative insight on how coders understand the task, and it can serve as proxy to assess the quality of the coding work.

The following experiment results are based on a task that contains several free-text fields. The specific task is insightful with regard to gaming detection using free-text fields. The task does not produce any trusted results (which are based on worker performance in the test questions). Therefore, the following presentation of the free-text replies are based on the untrusted results usually discarded by the platform, but available for inspection. In addition, any type of worker was allowed to work on the task (the crowdsourcing platform also offers to select skilled workers from a pool of workers who performed consistently well in previous tasks, i.e., high-performing workers).

The task is to assess texts as to whether they contain reflection categories. Each text is approximately 500 words long. The coders were asked to copy sentences from a given text and paste them into a free-text field as evidence of a specific reflection category.

For example, the coders were asked to find evidence of the item 'The writer recognises that something is not as it should be' in the following text:

> 'I want to reflect on the dance lesson with year 8, and in particular on the situation that arose with Ben, though I think that there are wider issues to be considered than just Ben. The situation left me feeling guilty and inadequate as a teacher. I see myself as having failed to prevent this situation and I suspect that none of us gained from it'.

The original text was obtained from Moon (2004), and shortened for the sake of illustration.

As an example, the workers could have copied the following sentences as evidence of the item and pasted them into the text field: 'The situation left me feeling guilty and inadequate as a teacher' and 'I see myself as having failed to prevent this situation and I suspect that none of us gained from it'.

In total, 1,569 results from 272 coders were gathered. Based on the text field replies of the coders, their answers were assigned to one of the following categories:

- **On-task replies:** Coder worked on the task as formulated in the task instruction (35 coders).
- **Copy:** Coder either repeated the same text chunk for all text fields, or copied the instructions, or copied segments of text without taking care of sentence beginnings or endings (101 coders).
- **Random text sequence:** Coder randomly pressed keys on the keyboard (46 coders). Some examples are: 'ghjghjghjgjghj', 'fadfda', or 'aaaaaaaaaaaaaaaaaaaasssssssssssssssssfffffffffffff'.
- **Unrelated:** Coder wrote text replies not based on the text of the task (31 coders). Some examples are: 'youtube marketminder forbes', 'yes i guess its ok', or 'Www.yahoo.com'.
- **Empty:** Coder did not copy-and-paste any sentences into the text field (17 coders).
- **Nonsense:** Coder provided nonsense answers using, for example, the 'lorem ipsum' filler text (21 coders). Some examples are non-English words such as 'lacrima lacrima lacrima', 'lorem ipsum dolor sit amet', or 'bloomys131082'.
- **Absent:** Coder rated everything as absent, i.e., no evidence of one of the reflection elements in the texts (12 coders).

- **Numbers:** Coder filled text field with numbers and not text (nine coders). Some examples are: '123 213', '12/12/1234', and '312321123'.

Table 6 summarises the average time in seconds dedicated to each page for each category.

There is a large difference in the time dedicated to each page between the coders who were on-task and those who were off-task. On-task coders required, on average, 348 seconds (SD = 527), whereas off-task coders required, on average, 53 seconds (SD = 81).

Category	Mean	SD	Pages
On-task	348	527	126
Copy	78	88	594
Absent	67	109	113
Unrelated	48	105	195
Numbers	34	22	53
Random	26	27	298
Empty	20	41	83
Nonsense	17	25	107

Table 6: Average time required per category

Completing the task according to the instructions required significantly more time than any of the gaming strategies.

The different gaming strategies result in an overall lower amount of time dedicated to each task, thus maximising the financial reward of the coding work.

The free-text answers can be used as proxy to determine which coders worked on the task, and which attempted to game the task. The text replies can serve as a mechanism to help understand coder motivation. Obvious off-task behaviour can be spotted efficiently by hand. In here, the text answer replies have been used to detect gamers. The completion time was not used as an indicator for gaming behaviour.

Similar to the test questions serving as quality measure, the open text responses can be used to automatically determine those coders attempting to game the task. The

results of the experiment show several repeating gaming strategies, for which an array of methods was developed to obstruct. These automated methods are described next.

5.6.2 *Custom validators*

The chosen crowdsourcing platform[11] provides several simple validators. A validator checks the entry of the coder according to several patterns. For example, text field validators check the minimum and maximum length of characters in an entry, whether the entry contains only text and not numbers, and whether the entry is a currency, email address, URL, or other regular expression-related validators. If the text reply from a coder does not match one of the patterns, a feedback message is displayed. The coder can only submit a task if all text replies conform to the validators.

These validators can help ensure that at least the free-text response is of a certain length, or that it is a text response. However, they have limitations for making a task robust against the observed gaming behaviour outlined in Section 5.6.1 'Gaming behaviour'.

The crowdsourcing platform allows the overwriting of existing validators with one's code. The client-side validator code is written in JavaScript, and it can be adjusted. In rare cases, the coders disable the JavaScript validation. For this case, the platform relies on server-side validation. For the latter scenario, custom validators can fail because it is not possible to modify the server-side validation.

A common gaming behaviour is to repeatedly copy the same text, usually text found on the web page that displays the task, into the free answer text fields. An algorithm that determines text similarity can help detect duplicated text and provide feedback to the coder that the answers are too similar to the web page text.

11 http://crowdflower.com

Several custom algorithms have been developed to alleviate the problem of repeatedly copying the same text into all text fields. These algorithms assume that on each page, several free-text fields are present, and that valid text responses are sufficiently different from each other. These prerequisites have to be considered during the task design process.

The implementation uses n-grams of word tokens from all open answer replies of a page in order to measure repeated occurrences of word sequences of a specific length.

In addition to checking the text of the free-text field, another algorithm checks whether any text on the task page was copied and then pasted into the free answer text field.

An algorithm based on a collection of common random keyboard patterns is used to combat random text sequences. In addition, a regular expression detects whether a character is repeated frequently (for example, 'aaaaa').

To ease the problem of nonsense answers, an algorithm that checks whether each reply contains at least one word of the most frequent English words is used. This helps minimise text replies in languages other than English, and texts full of nonsense words.

All custom validators provided a dialogue box that indicated to coders the problems with their text reply. These hints were also explained in the task instruction. In order to ensure the coder reads the instructions, another algorithm can be used to check whether all text replies start with a special symbol. If this symbol is not present, the dialogue refers the coder to the instructions.

These validators aim to make the task robust against some of the outlined gaming strategies. They make it more difficult to successfully game the system, and thus lower the reward for gaming. The assumption is that, if it is more difficult to game the system when actually performing a given task, the gamers will either stop the task, or focus on solving the task as instructed.

The automated detection of gaming strategies in texts using these validators has to maintain balance between making it difficult for gamers and not frustrating workers. These custom validators have been used for several smaller experiments, and based on the experience, they indeed make the task robust against gaming because they reduce the amount of off-task text replies. For example, in an experiment that made use of these custom validators, the manual inspection of the text field responses revealed that only eight out of 259 coders had to be removed because they provided off-task text responses. This is a considerable reduction compared to the 25% of workers that had to be removed by Schnoebelen and Kuperman (2010).

5.6.3 *Amount of ratings*

The manual method to assess reflection in writings is usually based on two to three trained coders (see Section 3.1.4 'Manual reflection detection performance'). Under the traditional premise of having at least three ratings per item, this practice is not recommended for crowdsourcing, especially when controlling the task for gamers. This is especially true for experiments with a smaller amount of assessed units. For example, in one experiment, a total of 32 sentences were rated. There was one coder who dedicated an extremely short amount of time to the task and was identified as a gamer. This one gamer was responsible for coding 16 sentences, which means that half the sentences received only two ratings. A higher amount of coders for each unit helps reduce the risk of not having sufficient ratings for each item.

There is no definitive answer to the question of what is the best amount of ratings requested from the platform for each unit. The research outlined in Section 5.5.1 'Research on crowdsourced annotation quality' reported five to 50 ratings per unit. The amount of coders necessary depends on the task, and should be determined with small pilot studies. The pilots that led to the task design for the

corpus generation process were based on approximately five ratings; after filtering for gamers, this resulted, in most cases, in at least three ratings per item. To determine the reliability of the process, more ratings are necessary, especially if reliability is determined with the split-half approach outlined in Section 5.7.4 'Reliability'. There, a sentence rating is determined through rating aggregation. The final vote to label sentences is based on the majority of approximately four to five ratings. Aggregating several ratings into one is an approach commonly used in crowdsourcing (see Section 5.5.3 'Aggregating crowdsourcing results'). This also means that in order to determine the reliability between two majority vote raters, at least twice as much ratings per item are necessary. Section 5.7.4 'Reliability' outlines this scenario (see p. 182). For example, in order to determine the reliability of two majority vote raters that based their vote on five ratings, ten ratings are necessary (five for each majority vote rater). However, it is not necessary to annotate all sentences in a dataset with ten ratings per sentence in order to determine reliability because such reliability can be determined on a subset of all units (see Section 4.2 'Evaluation criteria and metrics'). Once reliability is found to be sufficiently high, at least five ratings suffice to annotate the remaining units. The reason is that, once the process with the majority vote rater of five ratings is found reliable, this majority vote rater will likely produce the same reliable results in the future. Thus, for the actual task, five ratings suffice.

Here, approximately seven ratings are requested for each unit. The majority vote to assign a label to a unit is then based on seven, and not five, ratings. The additional two ratings were added in order to aid in the decision-making process as to exclude the unit from the dataset.

5.6.4 *Multiple-choice vs. rating scale*

Another experiment tested the use of multiple-choice questions. The intention is to reduce effort for the coders because the coder only has to choose one option, instead of coding several Likert scaled items. The design setup uses several test questions, a free-text field to gather coder responses, and custom validators to block obvious nonsense answers. Each page of the task contains four blocks of sentences. A coder could rate a maximum of 30 sentences. The qualification test consists of four questions randomly chosen from a large pool of test questions, and on every other page, another test question is mixed among the actual questions.

The coders were asked to select the best fitting category for the given sentence according to the following multiple-choice questions:

- □ Something should have been done differently
- □ Reasoning
- □ Takes another perspective into account
- □ Something was successfully achieved
- □ Something is interpreted in a new way
- □ None of these

The limitation of using multiple-choice questions over several rating scales is evident when filtering sentences with low overall agreement. These are sentences where the coder ratings indicated that they could not clearly place the sentence in one or another category. This might be the case for sentences that express two or more categories. Such sentences do not clearly indicate in which category they belong, and thus filtering them leaves sentences clearly identifiable as belonging to a category. Such are the sentences we wanted for building the corpus for the automated detection of reflection.

In total, 1,000 sentences received five ratings. Filtering reduced the amount of sentences drastically. For example, if only those sentences were maintained where three or more raters (out of five) agreed, the remaining sentences were 29 for the category something should have been done differently, 80 for reasoning, 19 for perspective, 80 for intention, 67 for achievement, 16 for interpreted in a new way, and 332 for none of these. From 1,000 sentences, 632 remained, out of which 291 express reflection categories.

If only those sentences were maintained where four or five raters agreed, the remaining sentences were 12 for something should have been done differently, 22 for reasoning, 4 for perspective, 42 for intention, 26 for achievement, 4 for interpreted in a new way, and 191 for none of these. From 1,000 sentences, 301 remained, out of which 110 express reflection categories.

Filtering resulted in only a small percentage of sentences remaining from the original 1,000, and those that remained were highly agreed. Using multiple-choice questions might be quicker when coding; however, the risk is that many sentences cannot be labelled with confidence. The difficulty is seen in the nature of the multiple-choice format in combination with the nature of the coding schema and the unit of analysis. Sentences can contain phrases, which can belong to several categories, and thus determining the best option might be difficult. Querying each question individually with Likert scales allows coders to rate sentences with multiple options allowing them to rate sentences regarding several categories of reflection.

The pilots with Likert scales did not lead to such drastic reduction of items after filtering. For example, in another experiment with 1,000 sentences where filtering was applied if less than eight out of ten people agreed on sentences, for the perspective category, 94 items were highly agreed as perspective sentences. In addition, Likert scales allow the consideration of highly agreed sentences from another dimension of

the scale. In this case, 414 sentences were highly agreed as not expressing a new perspective.

5.6.5 *General recommendations*

From the experiences observed during the pilots, several recommendations can be derived. The most important is to conduct several smaller pilots with crowdsourcing in order to gain an understanding of the design decisions that influence the quality of the results. Test questions are important because they help the workers to better understand the task, and it is a first barrier for filtering gamers. Test questions should be monitored closely and adjusted during pilots. Worker feedback is a valuable source of information for their understanding of the test questions.

Instead of administering a batch with many units, it is recommended to work with smaller batches. This allows assessment of the quality of each batch, reaction to worker comments, and the ban of obvious gamers from the next batch. For example, the experiments conducted with 1,000 sentences were divided into ten batches of 100 sentences each.

Gaming is a problem, especially if no restrictions with regard to the worker skill level are set. No restrictions means having access to the largest pool of workers, but also that effective mechanisms have to be in place for detecting gaming behaviour. In addition to time as an indicator of workers who complete each task page in an extremely short time (shorter than possible to work on the task diligently), free-text answer questions can be helpful to detect gamers. Time should not be considered as the sole basis for determining gaming behaviour because it might punish high performing workers. The custom validators outlined in Section 5.6.2 'Custom validators' considerably reduced the amount of gamers.

The amount of requested ratings per unit should be determined in pilots that help assess the amount of gaming, and with it, the number of annotations that have to be discarded after each task. Therefore, two to three ratings per unit, as in the traditional content analysis setting, is seen as too small. Here, five ratings were used to test different task designs, ten were requested to test the reliability of the task design, and seven to retrieve the annotations.

The use of multiple-choice items compared with separate rating scales runs the risk of having an extremely reduced set of results after filtering for highly agreed units. Therefore, the use of rating scales is recommended for the coding schema outlined in Section 5.7.2 'Task design'.

The experiences outlined from using the crowdsourcing platform were instructive for the task design used to generate the datasets for reflection detection.

5.7 ANNOTATION TASK

This section describes the annotation task used to generate the labelled datasets for the automated detection of reflection.

Section 5.7.1 'Task setup' describes the setup configuration of the crowdsourcing platform that has been used for all tasks. Section 5.7.2 'Task design' outlines the task design used to retrieve the ratings according to the model categories stated in Section 2.3.2 'Common reflection categories'. The experiences outlined in Section 5.6 'Summary of pilots' were implemented into the task design.

Section 5.7.3 'Participants' provides descriptive statistics of the participants of the task. The reliability of the coders is estimated in Section 5.7.4 'Reliability' and the validity of the model for reflection detection is measured in Section 5.7.5 'Validity'.

The result of the annotation task are datasets of annotated sentences with common reflection categories. Section 5.7.6 'Quality standard and datasets statistics' reports

descriptive statistics for each dataset after the application of the quality standard outlined in Section 4.6.1 'Dataset generation process'. These datasets are used in the evaluation part of this thesis.

5.7.1 *Task setup*

The general setup is the same for all tasks. The setup configuration considers the experiences observed during the pilots (see Section 5.6 'Summary of pilots'). A task on a crowdsourcing platform is a piece of the work distributed to the workers of the crowdsourcing platform. A task is comprised of an instruction and units. The instruction explains the task and provides an example of what is expected from the raters. Each unit contains a sentence and several items with which the workers had to rate each sentence. The items are operationalisations of the common reflection categories (see Section 2.3.2 'Common reflection categories'). Each item has to be rated on a six-point Likert scale, ranging from disagree to agree, and for the level of reflection item from descriptive to reflective. A free-text field is placed below the items, and it asks the raters to justify their choice of why they think the sentences are descriptive or reflective. The task description and unit design are fully described in Appendix E 'TASK DESIGN'.

The crowdsourcing platform[12] advertises the task to the workers, and those who decide to work on the task see it in the form of a web page. Each page contains the instruction and four units. The platform automatically assigns a sentence to a unit from a pool of sentences. The same sentence is never assigned twice to the same worker. After the worker completes all items and the text-field of a page, they can move to the next page, or leave the job. Every new worker undergoes a qualification test composed of four test questions that are units with known answers to the system.

12 CrowdFlower http://crowdflower.com

Such test questions are manually defined by the task requester, and they appear as every other unit, but in the case where the worker makes a mistake, the wrong answers are marked and the worker receives feedback with an explanation for the correct answer. The test questions serve two purposes. First, the feedback helps the worker better understand the task, and second, those workers who answer too many questions wrongly do not receive another page. If the worker performs well during the initial test, they are admitted to the next page, which contains the actual units to rate. Each of these successive pages has one unit out of four that is also a test question. The crowdsourcing platform keeps track of the performance of all workers of a task. If the performance falls below a threshold, the worker does not receive another page. The test questions are from a pool of 50 designed and tested by the researcher in initial pilots. The test questions are balanced, which means that every item has approximately the same amount of test questions for both the absence and presence of an item. On average, two items of a test unit are test questions. For each test item, a feedback answer is stored in the system explaining why the sentence expresses either the presence or absence of a category. In addition to the test questions, each task is equipped with the custom validators outlined in Section 5.6.2 'Custom validators'. Custom feedback is generated by the validator for the case where gaming behaviour is detected.

The platform provides another method for controlling worker quality: it offers the option of allowing into the task only those workers known for accurate work in previous tasks. Several levels can be set; the higher levels restrict the pool of workers, which also means that the task takes longer. For the data annotation tasks, we chose unrestricted access because previous pilots found that the test questions and validators are effective methods for controlling the annotation quality.

The platform allows the geographical distribution of the tasks. Because the task is in English, the choice fell on countries with English as the native language, for example, the United Kingdom, the United States, Australia, and Canada.

The entire dataset of sentences is first shuffled randomly. This ensures that during the annotation task, every rater receives a random set of sentences, thus avoiding a potential sequencing effect that can influence the rating decision. Because the crowdsourcing platform distributes these sentences to available workers, there can be a case where one worker sees all the sentences in their original order, whereas another worker receives not in their original order. The random shuffle of the sentences avoids this scenario. Subsequently, the dataset is divided into batches of 100 sentences, and each batch has the same task setup and task design. Each batch is assessed individually with regard to the quality of the annotations, the difficulty of the test questions, the response to gamers, and the question answers of the workers. Although the overall setup for the tasks does not change significantly, smaller adjustments are made to the platform settings during some of the tasks. For example, allocation of funds, adding or removing worker channels[13], or adjusting the test questions. Batching the tasks has the benefit that the experiences made in one batch can inform the next batch.

5.7.2 *Task design*

The same setup described in Section 5.7.1 'Task setup' is used for all batches of the entire task. The task setup describes the settings of the crowdsourcing platform. This section describes the choice of items with which the workers are asked to rate each

13 The crowdsourcing platform advertises the task to many crowdsourcing channels, each with its own workforce.

sentence. These items are indicators (operationalisations or proxies) of the common reflection categories (see Section 2.3.2 'Common reflection categories').

As outlined in the theory part of this thesis, Section 2.3 'Model for reflection detection' reflective writing has many facets that have been classified, and the result of this process is the common categories of reflective writing (see Section 2.3.2 'Common reflection categories'). They form the components of the theoretical model that guides this thesis. This section presents the chosen indicators for these common reflection categories. Each indicator represents the item used to code each sentence. Because the indicators are derived from the common reflection categories, they are proxies for these categories, and represent specific operationalisations of the common categories of reflective writing. Furthermore, these indicators can be understood as facets of a common category. This also means that they do not cover all imaginable facets of reflective writing, but they cover important facets.

This narrowing to specific indicators is similar to the approach taken by the research described in Section 2.2 'Models to analyse written reflection'. There, indicators were developed based on their soundness with regard to the chosen theoretical reflection framework. Section 2.2 'Models to analyse written reflection' provided an overview of all these indicators, and showed the variety of indicators used to measure reflection. Here, the guiding frame to develop reflection indicators are the common reflection categories derived from the models described in Section 2.2 'Models to analyse written reflection'.

Table 7 shows all indicators and their mapping to the common reflection categories. These indicators were tested in the smaller pilots that led to this task design (see the summary of these pilots in Section 5.6 'Summary of pilots'). The author reported on older versions of these indicators in Ullmann et al. (2012) and Ullmann et al. (2013).

Category	Indicator
Description of an experience	The writer describes an experience he or she had in the past
Feelings	The writer describes his or her feelings
Personal	The writer describes his or her beliefs
Critical stance	The writer recognises difficulties/problems
Perspective	The writer takes into account another perspective
Outcome	The writer intends to do something
	The writer has learned something
Reflection	The sentence is descriptive ... reflective

Table 7: Indicators of the common categories of reflective writing

The first two indicators for 'Description of an experience', and 'Feelings' are direct mappings from the category to their indicators. For the category 'Personal', an indicator is chosen that asked whether the sentence expresses a belief of the writer. As outlined in Section 2.3.2 'Common reflection categories', awareness of personal beliefs is an important part of reflection because these beliefs govern one's actions. The indicator for the category 'Critical stance' focusses especially on the awareness of problems. The identification of problems is the first step in a critical analysis: it is seen as the most important step because, without such awareness, the problem cannot be analysed. The indicator 'Perspective' asks for evidence of the consideration of other perspectives. The type of perspective is left open deliberately and not narrowed to specific perspectives, such as social context, historical context, theory, or other (see Section 2.3.2 'Common reflection categories'). Two operationalisations are chosen for the category 'Outcome' to cover the retrospective dimension of the 'Outcome' category by asking whether something was learned, and the prospective dimension that captures action intentions.

These indicators cover a wide range of aspects of written reflection. As such, they are good candidate indicators to test their potential with regard to their automated detection.

Furthermore, these indicators formed the unit of the task setup described in Section 5.7.1 'Task setup'. Each indicator formed one item. The following list shows the items for the task and their short names, which are used henceforth for easier referencing. The short names are written in bold font face next to each item.

- **Experience:** The writer describes an experience he or she had in the past.
- **Feelings:** The writer describes his or her feelings.
- **Beliefs:** The writer describes his or her beliefs.
- **Difficulties:** The writer recognises difficulties/problems.
- **Perspective:** The writer takes into account another perspective.
- **Intention:** The writer intends to do something.
- **Learning:** The writer has learned something.
- **Reflection:** The sentence is descriptive ... reflective

Each item is rated on a six-point Likert scale ranging from disagree (1) to agree (6), with the exception of the last item **Reflection**, where the item ranges from descriptive (1) to reflective (6). This is conform with the models of reflection that asses the depth of reflection, where the depth ranges from descriptive/non-reflective to reflective (see Section 2.2 'Models to analyse written reflection'). The reasons for choosing rating scales is outlined in Section 5.6.4 'Multiple-choice vs. rating scale'. The entire task design can be found in Appendix E 'TASK DESIGN'.

The task design is used for all 51 batches advertised to the raters over the crowdsourcing platform. From the 5,131 sentences from the sampled text collection (see Section 5.2 'Sampling text collection'), 50 sentences are used as the test questions, thus leaving 5,081 sentences for the annotation task.

5.7.3 *Participants*

In total, 3,133 people participated in the task, mostly from the United States and the United Kingdom. A total of 181 were identified as gamers. They were manually

detected based on their free-text responses (see also Section 5.6.1 'Gaming behaviour'). Ten ratings per sentence were requested for 1,300 sentences from the crowdsourcing platform. After initial inspection for reliability, which was found as sufficiently high, seven ratings per unit were ordered for the remaining 3,781 sentences (see Section 5.6.3 'Amount of ratings'). After removing all the gamers, an average of 7.92 ratings were retrieved for each sentence.

5.7.4 *Reliability*

Reliability is one of the main evaluation criteria, as outlined in Section 4.2 'Evaluation criteria and metrics'. This section outlines the considerations made to calculate reliability before presenting the results.

There are different ways for determining reliability. Here, the decisions made to calculate reliability are aligned to the requirements of the dataset for machine learning and the aggregation of ratings to determine labels.

The machine learning dataset should contain labelled sentences that indicate whether the reflection category is present or absent. This means that the coder ratings should be in a similar format that indicates presence or absence. Reliability could have been determined with all six levels of the Likert scale; however, in anticipation of the dataset format, the scale ratings are dichotomised. The first three levels (1, 2, and 3) are aggregated to indicate absence, and the last three levels (4, 5, and 6) to indicate presence. There are many ways for grouping these levels, and they can influence reliability (see Alonso and Mizzaro (2012, p. 1061)), but here, the simple divide in halves is chosen instead of optimising the reliability for specific division combinations.

As described in Section 5.5.1 'Research on crowdsourced annotation quality', the annotations of the crowdsourced workers achieve high reliability when aggregated

(see especially Section 5.5.3 'Aggregating crowdsourcing results'). This is a different approach compared with the manual content analysis of reflective writing (see Section 3.1.4 'Manual reflection detection performance'). There, the rationale is that, if the reliability determined on a sub-sample of the dataset is sufficiently high, a single coder can continue the coding process alone (examples can be found also in the literature cited in Section 3.1.4 'Manual reflection detection performance', for example Fischer et al. (2011, p. 168), or Poldner et al. (2014, p. 359). See also the discourse on reliability in Section 4.2 'Evaluation criteria and metrics'). In our case, the reliability for individual coders is, for most indicators, rather low, thus continuing the annotation task with only one coder is not justified (see Appendix D for the reliability values).

On the other hand, the crowdsourced coding process is based on the idea that the final determination on the label is derived from the collective decision of the crowd of raters. This is also the approach followed here. Therefore, it is important to determine reliability for this aggregated method and not for the individual level, as reported in the content analysis section. Section 5.5.3 'Aggregating crowdsourcing results' outlined several approaches to aggregate the crowdsourced results. A simple and commonly used strategy is to base the decision according to majority vote. Each rater votes as either presence or absence, and the one with the most votes is selected. Another way of aggregating the results would be to use all six levels of the rating scale in order to compare the average (or median) of the ratings of one group of coders with the average rating of another group of coders. However, the majority vote approach is more suitable because it casts voting in a binary format as needed for the dataset.

When using majority vote as the aggregation strategy, one has to consider how to manage 'ties'. A tie exists when the same amount of ratings indicate presence and absence for a reflection indicator. One way to manage this situation is to randomly assign a label to these cases. This strategy of tiebreaking at random has the benefit that

the reliability can be calculated over all cases because each case is assigned randomly with a label (see Snow et al. (2008, p. 258) described in Section 5.5.1 'Research on crowdsourced annotation quality'). However, this strategy is not deemed useful for the context of this study because the interest lies in a dataset where we are confident that the labels are assigned based on the majority of judgements, and not influenced by a random label generator. Therefore, the approach taken here is to treat ties as cases where the raters could not determine whether the sentence expresses the absence or presence of an indicator. Because no decisions could be made, they are removed from the dataset so that only those sentences where a decision is possible remain. This is similar to the approach used in the method of content analysis where the units that could not be assigned to a category are marked as 'unknown', or 'unrelated' (for example, see Chamoso and Cáceres (2009, p. 205) and Kovanovic et al. (2014)). The chosen approach has the benefit that all the labels in the final dataset are based on a majority vote, and not on randomly assigned labels.

In particular, in the context of crowdsourcing, it is common that not all units receive the same amount of ratings, for example, gamer ratings have to be removed. An extreme and rare case is when none or only one rating remains for a unit. However, a single rating would contradict the purpose of majority voting. Thus, a decision can be made as to when to discard a data point because of a lack of ratings. For the calculation of reliability in this chapter, a minimum of four ratings is considered as the minimum.

After this outline of the considerations for calculating reliability, the focus shifts to the presentation of the results. Two cases are considered.

The first case is based on simple majority voting, where the majority vote is determined for each sentence. For example, if six out of ten ratings agreed that a sentence is reflective, it is labelled as reflective. This approach that uses a simple

majority vote is common in the context of crowdsourcing (see Section 5.5.3 'Aggregating crowdsourcing results').

The second case takes the strength of majority voting further. As outlined in Section 4.6.1 'Dataset generation process', a more strict quality standard than simple majority vote can be applied to label the sentences. Stated simply, we trust a judgement more when more people agree on a unit. For example, if a sentence is rated by eight out of ten people as reflective, we are more certain about the sentence being reflective compared with a sentence where six out of ten people agree. A four-fifth majority is a more strict quality standard than simple majority. Because it is the aim to generate a high quality dataset in which all sentences receive substantial support to justify their labels, this more strict quality standard is used later to select the sentences for the final dataset. It ensures the validity of the dataset because only sentences that received substantial support of being reflective enter it. Here, we present the results of both cases. The reason for presenting the first case is that it is more common to use the simple majority vote, and thus it can be used to compare our results with the results of other researchers.

Agreement and reliability are determined by following the idea of randomly dividing the ratings for each sentence into two groups. Within each group, the label is determined by majority vote. Reliability is then calculated based on these two ratings, which measure the reliability of the two majority vote raters. The parallel to the classic form of determining reliability with two coders is that instead of having two coders, coding is determined here by two coding devices. The coding device is the majority voting system. The first coding device gathers five ratings from five different people and determines the final vote. The second device does the same with another set of five people. As in the traditional case, if the two coders differ in their ratings, the reliability is low. If they rate similarly, the reliability is high. Here, we check how similar the two majority vote raters vote.

For example, if four out of five ratings of group A agreed that a sentence is reflective, it is labelled as reflective. If four out of four ratings of group B agree that the same sentence is also reflective, there is an agreement of both groups of raters on this sentence. This is essentially a comparison of two (aggregated) raters, in which the rating of each rater is based on the majority vote of many ratings.

Before describing the results, it is worth explaining the two ICCs (see Section 4.2 'Evaluation criteria and metrics' for the description of all models). ICC(1,1) is the ICC value calculated from the majority votes of each group of raters. The first '1' in ICC(1,1) denotes the first ICC model, and the second '1' is the ICC for individual raters. The k in ICC(1,k) indicates that it is the average rater model. ICC(1,k) considers the averaged values of all k ratings (every sentence is rated a maximum of ten times; thus, this is ICC(1,10)), whereas ICC(1,1) calculates the individual reliability of our two aggregated raters. The k versions of ICC are commonly used to retrieve reliability values for averaged raters, and therefore, it is included in order to compare it with the majority vote approach. As outlined above, there are several ways of measuring the reliability of aggregated values. Here, the majority vote approach is more suitable because this is the mechanism that later determines the final dataset. Therefore, we report ICC at the individual level with ICC(1,1). ICC(1,k) allows the comparison of the reliability of the averaged approach with the chosen majority vote approach.

Further, ICC(1,k) can serve to validate the majority vote approach. We would expect that, although their values are different, the larger values for one measure also result in larger values in another measure; similarly for smaller values. The assumption is that the majority vote approach correlates positively with the established ICC(1,k). Here, the validity of the new method is tested against the established ICC(1,k) measure for reliability. A suitable measure of correlation to test this assumption is Pearson's product moment correlation. Both ICC(1,1) and ICC(1,k) are normality distributed according to the Shapiro-Wilk test. In addition, inspection of the dot plot

of both indicates that their relation is linear. This suggests the suitability of Pearson's product moment correlation. The correlation between ICCs in Table 8 is statistically highly significant ($r = 0.99$, $p < 0.001$). This strong correlation between both measures is an indicator of the validity of the majority vote rater approach.

Table 8 shows the agreement and reliability values of the majority vote raters for all chosen indicators of the common categories of reflective writing. N denotes the number of sentences after removing ties and sentences with insufficient ratings. The sentences are mostly from the 1,300 sentences that were rated up to ten times (see Section 5.7.3 'Participants'). The %-sign signifies per cent agreement. The measures reported here are described in Section 4.2 'Evaluation criteria and metrics'. Cohen's κ assumes that all data are rated by the same two raters. Arguably, they are rated by the same meta-rater, which is the majority vote meta-rater. On the other hand, Krippendorff's α, Gwet's AC_1, and ICC1 do not require fixed raters, and thus are a better fit for this scenario.

Indicator	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)	ICC(1,k)
Experience	1239	0.92	0.83	0.83	0.83	0.83	0.89
Feelings	1060	0.88	0.72	0.72	0.78	0.72	0.82
Beliefs	1026	0.82	0.65	0.65	0.65	0.65	0.78
Difficulties	1140	0.86	0.71	0.71	0.71	0.71	0.84
Perspective	994	0.80	0.47	0.47	0.67	0.47	0.65
Intention	1224	0.93	0.71	0.71	0.92	0.71	0.83
Learning	984	0.80	0.55	0.55	0.64	0.55	0.70
Reflection	1015	0.84	0.62	0.62	0.72	0.62	0.75

Table 8: Reliability of two simple majority vote raters over all indicators

The two majority vote raters agree on all indicators at over 80%. Agreement on the indicator **Experience** is in the top bracket of the reported per cent agreement in Section 3.1.4 'Manual reflection detection performance'. The per cent agreement of all

other indicators is the middle bracket between 80% and 89%. The per cent agreement of all indicators is in the range of the per cent agreement reported in the research on manual content analysis of reflective writing.

All indicators, but one, rate above the 0.5 Cohen's κ benchmark for exploratory research outlined in Section 4.2 'Evaluation criteria and metrics'). Most are substantial, and the indicator **Experience** is almost perfect. Two have moderate reliability (**Learning** and **Perspective**). The rating reliability of the indicators **Experience**, **Feelings**, **Difficulties**, and **Intention** is in the top range of the reported Cohen's κ values in Section 3.1.4 'Manual reflection detection performance'. They did, however, not reach perfect reliability as was reported in two of the papers. **Beliefs** and **Reflection** are in the middle range, **Learning** is in the lower range of reliability. All these indicators are in the range of what can be expected according to the reported reliability values of the manual content coding process of reflective writings. The indicator **Perspective** had the lowest κ and was below the reported values of the research of the manual analysis of reflective writing.

The indicators of the three papers that reported ICCs in Section 3.1.4 'Manual reflection detection performance' are in the 0.6 to 0.7 bracket, some below, and some above. The raters of the indicators **Experience**, **Feelings**, **Difficulties**, and **Intention** exceed that bracket (see the column ICC(1,1). **Beliefs** and **Reflection** are in this reported range. The proxy **Learning** is rated within the bracket from 0.5 to 0.6. **Perspective** is below 0.5. It is notable that the group-based ICC(1,1) values are consistently lower than the ICC(1,k) reliability for average ratings. The split-half approach produces smaller reliability values compared with the ICC(1,k) model that considers all ratings.

According to Gwet's AC_1, the rater reliability for the indicators **Experience** and **Intention** are almost perfect. All others are substantial.

Thus far, we have discussed the reliability of the raters using the simple majority vote approach. Table 9 shows the reliability for the two four-fifth majority vote raters. The reliability of all, but one, indicators are almost perfect. The exception is the indicator **Perspective**, where the raters achieve substantial reliability. The four-fifth majority vote rater produced highly reliable results. The agreement and the reliability is in the highest bracket reported by the research of the manual content analysis of reflective writings (see Section 3.1.4 'Manual reflection detection performance').

Indicator	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
Experience	742	0.98	0.96	0.96	0.96	0.96
Feelings	514	0.98	0.94	0.94	0.96	0.94
Beliefs	376	0.97	0.93	0.93	0.94	0.93
Difficulties	540	0.97	0.94	0.94	0.94	0.94
Perspective	364	0.95	0.78	0.78	0.93	0.78
Intention	785	0.99	0.93	0.93	0.99	0.93
Learning	347	0.96	0.91	0.91	0.93	0.91
Reflection	413	0.97	0.92	0.92	0.96	0.92

Table 9: Reliability of four-fifth majority vote raters over all indicators

This concludes the part on the reliability of the reflection indicators. Two strategies were followed. The first explored the reliability of the simple majority vote raters in order to retrieve reliability values comparable with the reported reliability values from research on the analysis of reflective writings (see Section 3.1.4 'Manual reflection detection performance'). Similarly to the process described there, the majority vote raters rated the sentences according to the reflection categories. Overall, the reliability is in the range of the reliability values outlined there. In order to achieve these results, the ratings of approximately five raters were aggregated with the simple majority vote that determined the final label.

The second strategy determines the reliability of the four-fifth majority raters. This is important because the four-fifth majority vote will be applied to all sentences in order to determine a high quality final datasets. Naturally, the two four-fifth majority raters achieve higher reliability. Mostly, reliability is almost perfect.

After discussing the reliability of the data generation process, the focus now shifts to the second evaluation criteria outlined in Section 4.2 'Evaluation criteria and metrics', namely validity.

5.7.5 *Validity*

The chosen indicators are derived from the common reflection categories (see Section 5.7.2 'Task design'). Thus far, their validity (see Section 4.2 'Evaluation criteria and metrics') is based on the fact that they are sound derivations from the common categories on reflective writing. It is plausible to believe that they capture important facets of the common categories. This type of validity is called face validity (Krippendorff, 2012, p. 329). Most of the research outlined in Section 2.2 'Models to analyse written reflection' based their choice of indicators to assess the writings on their soundness in the context of the theory of reflection. With this approach, the indicators have face validity. Another way to determine the validity of the indicators is to collect empirical evidence with regard to their validity. However, the research examined in Section 2.2 'Models to analyse written reflection' only contained few reports on empirical evidence of validity. This observation resonates with the findings of Poldner et al. (2014, p. 31 f.). They noted that 'None of the 18 articles used methods like correlation analysis, group differences, or experimental or instructional interventions to gather evidence of for validity.' (Poldner et al., 2012, p. 31 f.). An exception to this is, for example, the study of Poldner et al. (2014, p. 363). They

correlated reflection levels to external criteria that was an assessment test. Using correlations is a common way for determining empirical validity.

The chosen task design (see Section 5.7.2 'Task design') makes it possible to empirically investigate the relationship between the indicators of the common reflection categories and the indicator that measures the two reflection levels, namely, descriptive and reflective. The assumption is that the breadth indicators of reflective writing are related to the level indicator **Reflection** (see Section 2.2 'Models to analyse written reflection' and Section 3.1.3 'Relationship between the descriptive and level reflection quality'). This section reports the results of the association between the reflection indicator and all indicators of the common categories of reflective writing.

The expectation is that all common categories indicators correlate positively with reflection. We would expect that, for example, sentences that express beliefs are more likely to also express reflection. It would speak against the validity of the indicator if the sentences that express beliefs were mostly rated as descriptive/non-reflective.

There are several correlation coefficients. Frequently, Pearson's product moment correlation is used to measure the correlation between two variables. However, this measure relies on several test assumptions, for example, the variables should be normally distributed and the relationship should be linear. A test of normality shows that a nonparametric correlation coefficient is more suitable. Spearman's rank correlation (ρ) is a nonparametric measure (Siegel, 1957; Brown and Hayden, 1985). Although binary data could have been used similar to Section 5.7.4 'Reliability', here, the Likert scale ratings are averaged in order to better approximate the rank correlation compared with binary ranks. In order to maintain this analysis similar to the process that determined reliability, the simple majority vote approach is used to include only those sentences on which a majority could agree[14]. The ranks are calculated based on their averages. For example, a sentence rated on average as 1.3

14 The differences of the rank correlations between this approach and an approach that uses all sentences were small (maximum 0.03 difference).

(descriptive) receives a lower rank than a sentence rated as reflective with an average of 5.3.

Spearman's ρ ranges between -1 and +1. A value of -1 indicates a negative rank correlation. This is the case when, for example, rising values of **Beliefs** appear with decreasing values of **Reflection**. A positive rank correlation indicates the expected relationship between the indicators of the common categories and the reflection indicator.

N signifies the amount of rated sentences (excluding sentences without a majority vote (i.e., 50:50) and sentences that received fewer than four ratings). The letter p represents the p-value.

Table 10 shows Spearman's ρ between **Reflection** and all indicators of the common reflection categories.

Indicator	N	Spearman's ρ	p
Experience	4659	0.50	0.00
Feelings	4628	0.76	0.00
Beliefs	4576	0.68	0.00
Difficulties	4570	0.48	0.00
Perspective	4505	0.32	0.00
Intention	4698	0.28	0.00
Learning	4526	0.54	0.00

Table 10: Correlation between reflection indicator and indicators for common categories of reflective writing

All rank correlations are statistically highly significant, but this is not surprising considering the large sample size. The direction of the rank correlation is as expected. This means that all indicators of the common categories positively associate with reflection. **Feelings** correlates strongly with **Reflection**. **Experience**, **Beliefs**, **Difficulties**, and **Learning** correlate moderately, and both **Perspective** and **Intention** are weakly associated with the indicator **Reflection**.

It is notable that the associations between **Reflection** and both **Intention** and **Perspective** are weak. The first indicator is a forward-looking concept that asks what

to do next. This is not a concept of looking back at past experiences or lessons learned, which is more commonly associated with reflection (see the Latin root of reflection outlined in Chapter 1 'INTRODUCTION') and can be seen in the moderate correlation with **Experience** and **Learning**. The second (Perspective) shifts the focus away from the thought sphere of oneself to other perspectives. It seems that **Reflection** is more associated with one's perspective than with the consideration of other perspectives. This can be seen in the moderate association with **Beliefs**, which summarises expressions of one's beliefs and perspective on the world.

The empirical evidence for validity in Table 10 corroborates the face validity of the indicators. The theoretically derived relationships between reflection and the common categories of reflective writings are also found in their rank correlations.

5.7.6 *Quality standard and datasets statistics*

The process described thus far enriched the dataset of 5,081 sentences with ratings for all indicators of the reflection categories. Each sentence was rated on average 7.92 times. This section outlines the standard that was applied to all datasets to ensure their quality and presents descriptive statistics of the final datasets. Section 5.7.4 'Reliability' shows that the strategy for determining the final label for each sentence by majority vote is reliable. Still, the majority vote approach can err, especially if the majority is thin. As outlined in Section 4.6.1 'Dataset generation process', one of the requirements for a suitable dataset for machine learning is that it contains a sizeable dataset of examples that *represent* the construct. The four-fifth majority vote is applied in order to generate these high quality datasets for machine learning.

A four-fifth majority vote requires that most ratings reflect the annotation. The more ratings agree on a label, the higher is the confidence that the sentence actually expresses that label. The higher the effort that has to be made to justify an annotation is at the

cost of fewer sentences available in the dataset, but these sentences are highly likely to represent their indicators. Other standards can be applied, but the four-fifth majority vote allows, for all indicators, sufficient data to train and test the machine learning models. The sample size is similar to the sample sizes of the related machine learning approaches (see Section 3.2.3 'Machine learning approaches'). The reliability of the four-fifth majority rater was estimated in Section 5.7.4 'Reliability' (see Table 9) and found sufficiently high. The per-cent agreement and Cohen's κ values are in the top bracket of the reliability reported in the research about the manual analysis of reflective writing (see Section 3.1.4 'Manual reflection detection performance').

The following Table 11 shows the amount of sentences that remain in the dataset when applying the four-fifth majority standard, and the average word count of all sentences[15]. Both are divided into two categories: absent and present. Present means that the sentences are highly agreed to express the indicator. Absent means that the sentences are highly agreed to not express the indicator. For example, 588 sentences are highly agreed to be reflective, and 1,695 are highly agreed as being descriptive/non-reflective.

	Number of sentences		Average of unigrams	
Indicator	absent	present	absent	present
Experience	1768	1505	20.77	24.46
Feelings	1820	786	20.31	25.81
Beliefs	1095	1159	18.78	25.45
Difficulties	1291	1340	17.26	27.00
Perspective	1678	307	17.45	29.32
Intention	3295	341	21.67	24.90
Learning	1153	679	18.13	26.70
Reflection	1695	588	20.47	25.46

Table 11: Statistics for annotated dataset

15 The number of sentences is higher than the number of sentences reported in the reliability Table 9 on page 186. The reason is that there at least 8 ratings were necessary to determine the reliability of the two four-fifth majority vote raters. This reduced the amount of sentences to the ones that had more than eight ratings. Here, the minimum of required sentences was four, which means that more sentences are available.

It is notable that most indicator datasets have more absent sentences compared with sentences that indicate their presence. The exception are the indicators **Beliefs** and **Difficulties**. Some indicator datasets are more equally balanced, such as **Experience**, **Beliefs**, and **Difficulties**, whereas other indicator datasets are more imbalanced. Such dataset imbalance is discussed again in the following evaluation part. As described in Section 4.6.2 'Research design', imbalanced datasets might influence the performance of the classifier. As outlined there, an oversampling strategy is used to evaluate the degree to which the imbalance of datasets is problematic for the selected machine learning algorithms.

The average count of unigrams shows that those sentences that are examples of the indicator have more unigrams than those sentences that do not show the indicator characteristics. Overall reflective sentences are longer than non-reflective sentences. For comparison, a blog has on average 13.2 words per sentence (Herring et al., 2004, p. 9). Several sample sentences for all the indicators can be found in Appendix F 'EXAMPLES OF THE DATASETS'.

5.7.7 *Summary*

The data generation process outlined in Section 4.6.1 'Dataset generation process' was implemented in this chapter. The process started in Section 5.1 'Identification of text collection' with a large text collection of academic writings, and ended in Section 5.7.6 'Quality standard and datasets statistics' with eight datasets annotated with regard to operationalisations of the common reflection categories (see Section 2.3.2 'Common reflection categories').

As described in Section 5.1 'Identification of text collection, the choice for text collection fell on the BAWE corpus. This corpus contains a large text collection of academic writings, which made it suitable as a data source candidate. Corpus

inspection showed that it indeed contains reflective writings, but also that such writings are less frequent. A sampling approach was devised to help determine a sub-sample of the corpus of personal academic writings (see Section 5.2 'Sampling text collection'). The outcome of this process was a collection of texts that were then automatically divided into sentences (Section 5.3 'Unitising text collection'). Several types of the analysis unit were considered. Sentences, as the unit, were used in the research of the analysis of written reflection, and were found useful for this data generation process. The dataset of sentences was then annotated in order to generate a labelled dataset necessary for supervised machine learning. Crowdsourcing was identified as the potential method for annotating this large text dataset of unlabelled sentences. The research outlined in Section 5.5.1 'Research on crowdsourced annotation quality' suggested that crowdsourcing is a viable option if the task design is properly specified (Section 5.5.2 'Research on crowdsourcing task design'), and if the ratings of several raters are aggregated (Section 5.5.3 'Aggregating crowdsourcing results'). Several pilots preceded the actual annotation task in order to specify the task design applicable to the context of the annotation of written text with regard to the reflection categories. Section 5.6 'Summary of pilots' summarised the pilot experiences. Especially important was the analysis of gaming behaviour in Section 5.6.1 'Gaming behaviour' that led to the development of custom text validators in order to reduce the amount of workers gaming the system (see Section 5.6.2 'Custom validators'). The pilots also helped approximate the amount of ratings necessary for inferring the label of the sentences (see Section 5.6.3 'Amount of ratings') and the choice of rating scale (see Section 5.6.4 'Multiple-choice vs. rating scale'). The experience of the pilots were summarised in the form of recommendations in Section 5.6.5 'General recommendations'. The experiences observed during the pilots informed the annotation task (see Section 5.7 'Annotation task'). The configuration of the task was described in Section 5.7.1 'Task setup'. The

task design outlined the choice of indicators used to rate the sentences according to the common reflection categories (see Section 5.7.2 'Task design'). The participants of the annotation task (Section 5.7.3 'Participants') rated all the sentences. The reliability of the rating process was confirmed in Section 5.7.4 'Reliability'. There, it was shown that the aggregation of several ratings in order to derive the underlying annotation has comparable reliability to the values reported in Section 3.1.4 'Manual reflection detection performance'. The validity of the indicators was empirically supported in Section 5.7.5 'Validity'. These results suggested that the data generation process is reliable, and the indicators of the descriptive common reflection categories are positively correlated with the level of reflection indicator. In addition to reliability and validity, the task design was the same for all workers and was administered with the same task setup, which aids with the objectivity of this process (see Section 4.2 'Evaluation criteria and metrics'). Section 5.7.6 'Quality standard and datasets statistics' outlined descriptive statistics of the final datasets, and explained the process taken to ensure that the only sentences entered in the datasets are highly agreed. This quality standard ensures the high quality of the datasets.

The datasets generated in this chapter serve as training and test data for Chapter 6 'EVALUATION'.

6

EVALUATION

This chapter presents the evaluation results of the machine learning algorithms on the problem of reflection detection, and it is structured according to the two main research questions (see Section 1.1 'Research questions'). Section 6.1 'Reflection' presents the results that answer the first research question, and Section 6.2 'Common categories of reflective writing' the results for the second research question. The research design for the evaluation can be found in Section 4.6.2 'Research design' (for a schematic overview see Figure 3 and Figure 4).

The first research question is: **Q1: Can machine learning algorithms be used to distinguish between descriptive and reflective text segments?** In order to answer this research question, three lines of investigation are proposed. Each of these lines focusses on a specific type of machine learning algorithm.

The first line of investigation is: **I1: Can tree-based machine learning algorithms detect the difference between descriptive and reflective texts segments?** Section 4.5.1 'Tree-based models' outlines the concrete implementation of the machine learning algorithms used to evaluate the dataset generated from the indicator **Reflection**.

The second line of investigation is **I2: Can rule-based machine learning algorithms detect the difference between descriptive and reflective text segments?** The algorithms used to evaluate their performance on a dataset of reflective and non-reflective/descriptive sentences can be found in Section 4.5.2 'Rule-based models'.

The third and last line of investigation is: **I3: Can high performance machine learning algorithms detect the difference between descriptive and reflective text segments?** Section 4.5.3 'High performance models' provides the details of the selected machine learning algorithms for this line of investigation.

The rationale for the three lines of investigation are described in Section 4.5 'Machine learning algorithms'. The discourse on choosing machine learning over other automated methods outlined in Section 3.2 'Related automated methods' can be found in Section 4.1 'General methodological considerations'.

All the machine learning algorithms for the three lines of investigation are trained on the same set of reflective and descriptive sentences and evaluated on separate sets of test data. The process outlined in Section 4.6.2 'Research design' was enforced in order to maintain the conditions equal for all classifiers and evaluate their differences. This concerns the aspects of data pre-processing, feature selection, and feature construction, as well as the resampling strategy, model tuning, and model selection tested in the model assessment step.

The second research question is: **Q2: Can machine learning algorithms be used to detect common categories of reflective writing?** Here, the high performance classifiers outlined in Section 4.5.3 'High performance models', and already tested on the dataset of reflective sentences for the third line of investigation, are used to evaluate their performance on the datasets generated from each indicator of the common categories of reflection. These datasets were generated by applying the same quality standard to all (see Section 5.7.6 'Quality standard and datasets statistics'). The reliability of all datasets is estimated in Section 5.7.4 'Reliability' (see Table 9). As with the first research question, all algorithms are trained and tested with the same set-up.

The following sections present the results. The discussion of the findings for the first research question can be found in Section 6.1.4 'Discussion of the results of the three

lines of investigation', and for the second research question, in Section 6.2.7 'Discussion of the results of the common categories of reflection'.

6.1 REFLECTION

This section reports the results of the machine learning algorithms on the problem of detecting sentences that are reflective and those that are descriptive/non-reflective. A summary of the performance of the machine learning algorithms for all three lines of investigation can be found in Section 6.1.4 'Discussion of the results of the three lines of investigation'.

Table 12 presents statistics for the training and test dataset derived from the ratings of the indicator **Reflection**. Some examples of the sentences from this dataset can be found in Table 47 in Appendix F 'EXAMPLES OF THE DATASETS'. As outlined in Section 4.6.2 'Research design', the sentences are pre-processed and their features extracted. The features space consists of 769 stemmed unigrams.

In total, 2,283 instances[1] form the dataset. These are divided into two parts: one used for model training, and the other reserved for testing (see Section 4.6.2 'Research design'). Each instance represents one of two classes, positive or negative, where positive means that the instance is reflective, and negative that it is descriptive. More instances are negative than positive. The oversampled dataset is generated by randomly selecting instances of the minority class until both instances have the same amount of instances. The oversampled dataset consists of 1,356 positive and 1,356 negative instances. The test set is not oversampled in order to assess the performance of the classifiers on the original distribution of instances. In total, 456 instances form the test set.

1 The word 'instance' is commonly used in machine learning to refer to an example or record; here, it corresponds to a sentence.

Statistics	Count
Number of features	769
Number of instances	2283
Number of training instances	1827
Number of positive training instances	471
Number of negative training instances	1356
Number of test instances	456
Number of positive test instances	117
Number of negative test instances	339

Table 12: Statistics about the training and test set

The training and test sets are the same for all the machine learning algorithms used in the following three lines of investigation.

6.1.1 *Results of the tree-based models*

In this section, the performance measures of the tree-based models are reported. As outlined in Section 4.6.2 'Research design', the models are trained on the training dataset. In order to find the best model candidate for each algorithm, the k-fold cross-validation resampling strategy is used and the model with the highest Cohen's κ is selected as the final model of the training phase. The test dataset is used to determine model performance.

Table 13 shows the results of the tree-based models. Five models perform equally high with a Cohen's κ of 0.64. These are the Classification and Regression Trees (CART) tuned over the cost and the maxdepth parameters, Conditional Inference Tree tuned over the mincriterion, and C5.0 Decision Tree trained on the oversampled training dataset. In most cases, oversampling leads to a reduction in reliability, with the exception of the C5.0 model. The highest achieved accuracy (per cent agreement) is 87%, which means that these models err on this test dataset for 13% of the cases. Krippendorff's α and ICC(1,1) are extremely similar to Cohen's κ. Gwet's AC_1 is

generally higher than the other reliability measures. As outlined in Section 4.2 'Evaluation criteria and metrics', the AC_1 is a reliability measure developed to better reflect reliability in imbalanced datasets. Here, the higher AC_1 values indicate that there is substantial reliability between the predictions of the tree-based models and the highly agreed sentences of the human raters. The next table demonstrates the agreement and disagreement details by inspecting the sensitivity and specificity of the models.

Method	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
CART Tree cost	456	0.87	0.64	0.64	0.79	0.64
CART Tree cost oversampled	456	0.84	0.59	0.59	0.73	0.59
CART Tree max depth	456	0.87	0.64	0.64	0.79	0.64
CART Tree max depth oversampled	456	0.80	0.55	0.54	0.65	0.55
Conditional Inference Tree mincriterion	456	0.87	0.64	0.64	0.79	0.64
Conditional Inference Tree mincriterion oversampled	456	0.85	0.62	0.62	0.75	0.62
Conditional Inference Tree maxdepth	456	0.86	0.61	0.61	0.78	0.61
Conditional Inference Tree maxdepth oversampled	456	0.81	0.56	0.56	0.67	0.56
C5.0 Decision Tree	456	0.85	0.61	0.61	0.76	0.61
C5.0 Decision Tree oversampled	456	0.86	0.64	0.64	0.78	0.64
C4.5-like Tree	456	0.85	0.60	0.61	0.77	0.61
C4.5-like Tree oversampled	456	0.84	0.57	0.57	0.74	0.57

Table 13: Reliability of tree-based models

Table 14 shows the additional performance measures often used to describe the performance of machine learning models (see Section 4.2 'Evaluation criteria and metrics'). These are the Area Under the ROC curve, sensitivity (Sens.), and specificity (Spec.).

Method	N	AUC	Sens.	Spec.
CART Tree cost	456	0.87	0.72	0.92
CART Tree cost oversampled	456	0.89	0.74	0.87
CART Tree maxdepth	456	0.87	0.72	0.92
CART Tree maxdepth oversampled	456	0.81	0.85	0.79
Conditional Inference Tree mincriterion	456	0.87	0.72	0.92
Conditional Inference Tree mincriterion oversampled	456	0.86	0.75	0.88
Conditional Inference Tree maxdepth	456	0.86	0.63	0.94
Conditional Inference Tree maxdepth oversampled	456	0.86	0.84	0.80
C5.0 Decision Tree	456	0.87	0.70	0.91
C5.0 Decision Tree oversampled	456	0.90	0.74	0.91
C4.5-like Tree	456	0.85	0.68	0.91
C4.5-like Tree oversampled	456	0.79	0.66	0.90

Table 14: Performance measures of tree-based models

All models have lower sensitivity than specificity. This is not unusual for models trained on datasets that contain more negative than positive instances. The algorithms had more examples of descriptive sentences, and thus were able to better predict them.

The oversampled CART tree tuned over the maxdepth parameter achieves a sensitivity of 0.85. This means that 15% of the reflective sentences are incorrectly predicted as descriptive. The model with the highest specificity at 0.94 is the Conditional Inference Tree tuned over the maxdepth parameter. A total of 6% of the descriptive sentences are incorrectly predicted as reflective.

Sensitivity can be exchanged for specificity, and vice versa, by defining different probability cut-off points. The default threshold to classify a sentence as reflective is a predicted probability of more than 50%. If the probability is lower, the sentence is classified as descriptive. A threshold different from 0.5 can be selected to raise specificity on the costs of sensitivity, or vice versa. A tool for visualising this relationship is the ROC curve. Figure 8 shows the ROC curve for three selected

models with a high AUC. These are the oversampled C5.0 Decision Tree, oversampled CART Tree (cost), and Conditional Inference Tree (mincriterion).

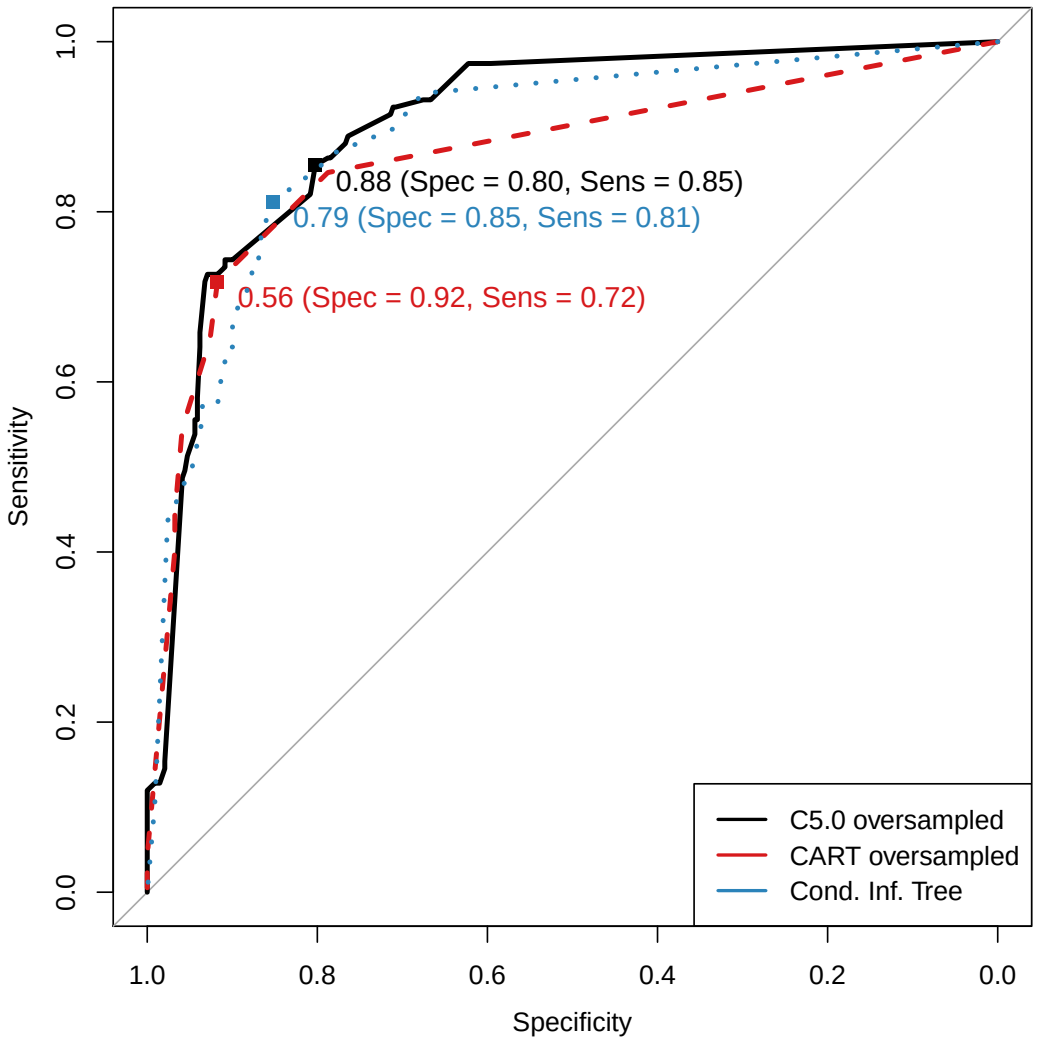

Figure 8: ROC-curves of tree models

As explained in Section 4.2 'Evaluation criteria and metrics', the diagonal line indicates the performance of a random model. The more area the ROC curve covers towards the left-upper area, the better is the performance of the model. A perfect model has a sensitivity of 1.0 and a specificity of 1.0.

Figure 8 marks the cut-off points that represent the highest sum of sensitivity and specificity. For example, the sensitivity and specificity of the oversampled C5.0 model for a 0.5 cut-off point is 0.74 and 0.91, respectively (see Table 14). If a cut-off of 0.88

were chosen instead of the default 0.5 threshold, the sensitivity would be higher (0.85) at the cost of specificity (0.80) (see Figure 8). Higher sensitivity is at the cost of lower specificity, and vice versa. The ROC curve shows that sensitivity can be increased without overly decreasing specificity to the point where the curve bends sharply to the right. Cut-off points in this area severely decrease specificity without much gain in sensitivity. All three models show this sharp bend, thus indicating limits in trading specificity for sensitivity.

One of the benefits of tree models is that they are highly interpretative because they form decision trees. Figure 9 shows the decision tree for the conditional inference tree model tuned over the mincriterion.

At the top is the root node with the most important feature. Here, this is the unigram 'i'. At the bottom, the leaf nodes are displayed as boxes that show the amount of sentences that satisfy all the decisions made on the path from the root node to the final decision node. The box's black bar shows the percentage of sentences that are negative (descriptive), and the grey bar shows the amount of sentences that are positive (reflective).

For example, a sentence that contains the unigram 'i' and 'that' are sorted into leaf node 17. From the training dataset, 239 sentences are sorted into this node. From these, fewer than 20% are labelled as descriptive (neg), and more than 80% are labelled as reflective (pos). A sentence without 'i' or 'me' are sorted into node number 3, which contains mostly descriptive sentences.

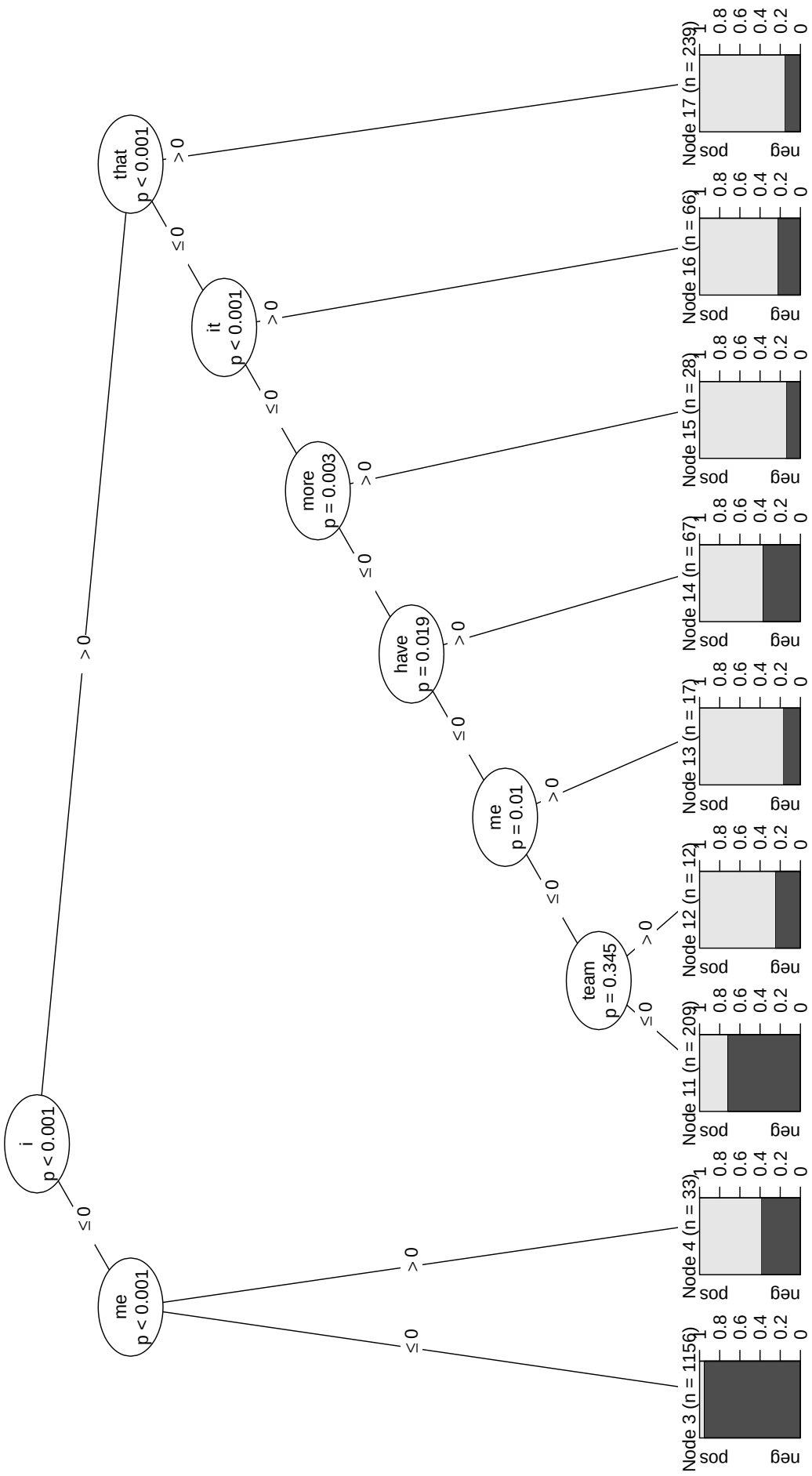

Figure 9: Tree visualisation of the Conditional Inference Tree

Ideally, the leaf node would only contain instances that are either positive or negative, but not both. However, the current tree nodes represent the best configuration for classifying unseen data as determined by the resampling strategy. More nodes would separate the sentences more clearly, but the nodes would then be more specific to the training dataset, and therefore, would not perform well on other datasets.

The discussion of these results continues in Section 6.1.4 'Discussion of the results of the three lines of investigation'. The next section presents the results of the rule-based models.

6.1.2 *Results of the rule-based models*

The rule-based models are trained and tested with the same data outlined in Section 6.1 'Reflection'. The following Table 15 shows the reliability of the rule-based models. A description for all rule-based machine learning algorithms can be found in Section 4.5.2 'Rule-based models'.

The C5.0 Rule implementation trained on the oversampled dataset achieves the highest Cohen's κ of 0.65, followed by the JRip algorithms with a reliability of 0.63. The simplest of the algorithm, OneR, can reach a per cent agreement of 79% and a Cohen's κ of 0.51 with only a single rule. The chance corrected AC_1 shows that 24% of difference is between the OneR and the C5.0 Rule trained on the oversampled data. The results of the OneR model can be seen as the baseline for this dataset. Any other algorithm should have higher reliability than 0.51.

The training of the models on the oversampled dataset is only beneficial for the C5.0 Rule algorithm. Reliability of the PART algorithm remains unaffected from oversampling, whereas the JRip algorithm suffers from oversampling.

Method	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
C5.0 Rules	456	0.84	0.59	0.59	0.74	0.59
C5.0 Rules oversampled	456	0.87	0.65	0.65	0.78	0.65
PART Rules	456	0.85	0.58	0.58	0.76	0.58
PART Rules oversampled	456	0.84	0.58	0.58	0.75	0.58
JRip	456	0.86	0.63	0.63	0.78	0.63
JRip oversampled	456	0.83	0.58	0.58	0.71	0.58
OneR	456	0.79	0.51	0.51	0.64	0.51

Table 15: Reliability of rule-based models

Table 16 shows the AUC, sensitivity, and specificity of the models. Here, oversampling has a positive effect on sensitivity. The oversampled JRip algorithm has the highest sensitivity, but also one of the lowest specificity. The PART rule algorithm has the highest specificity, but also the lowest sensitivity.

Method	N	AUC	Sens.	Spec.
C5.0 Rules	456	0.85	0.70	0.89
C5.0 Rules oversampled	456	0.85	0.74	0.91
PART Rules	456	0.83	0.62	0.93
PART Rules oversampled	456	0.79	0.66	0.91
JRip	456	0.81	0.68	0.92
JRip oversampled	456	0.85	0.79	0.84
OneR	456	0.79	0.78	0.80

Table 16: Performance measures of rule-based models

In order to investigate the relationship between the sensitivity and specificity of these rule-based algorithms, the ROC curves of the oversampled C5.0 Rule algorithm, PART model, and oversampled JRip algorithm are shown in Figure 10.

Compared with Figure 8, the dots on the ROC curve represent the standard 50% probability threshold. The ROC curve of the PART algorithm indicates that with a different probability threshold, sensitivity can be increased to the point where such increase would cause a steep specificity decrease. This is also true for the other two

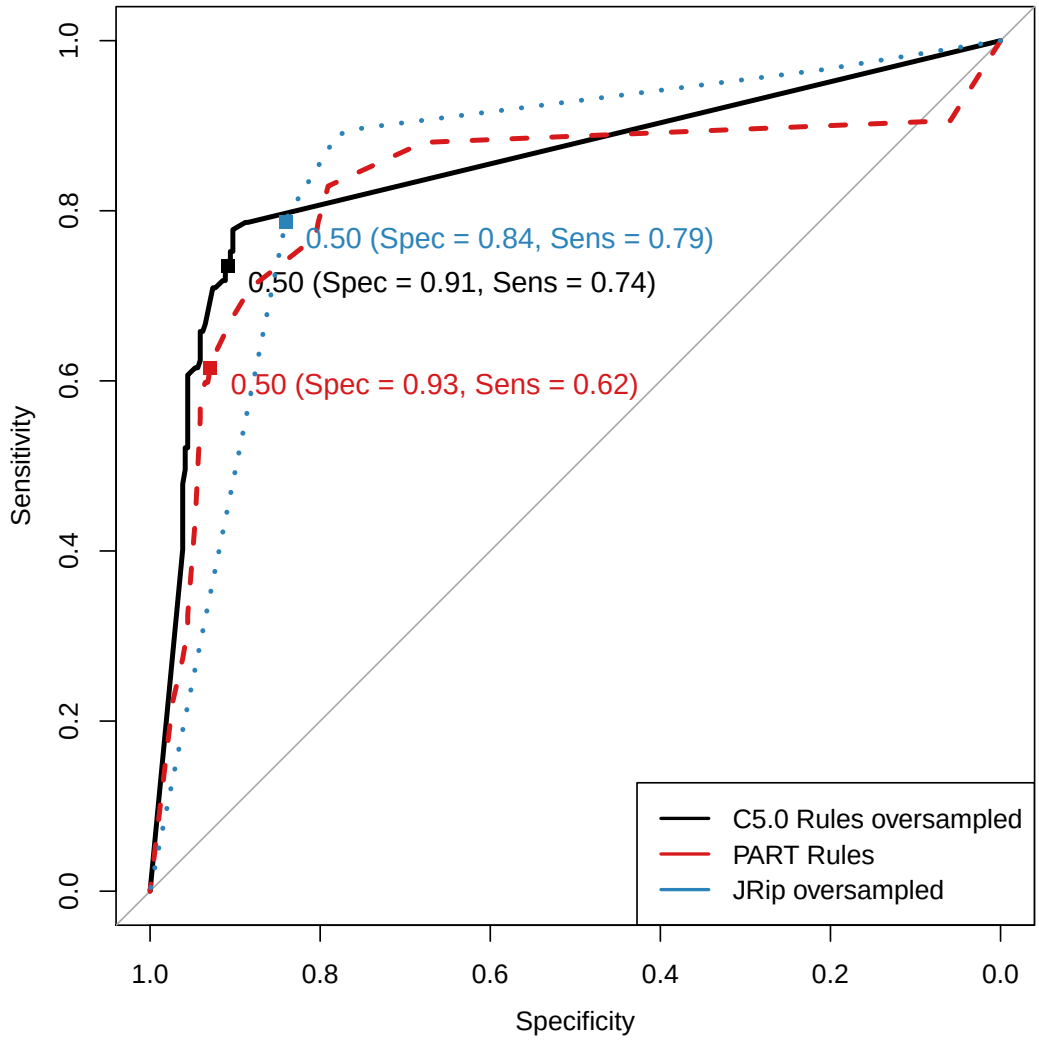

Figure 10: ROC-curves of tree models

models. Furthermore, none of the models can reach perfect sensitivity without zero specificity.

As with the tree-based models, the rule-based models can be inspected. Instead of a tree representation, the rule-based models follow the common notion of premise and conclusion.

As mentioned above, the simple OneR rule can achieve a Cohen's κ of 0.51 with a single rule. This rule is shown in the following listing, and it is based on the single unigram 'i' used to form the rule. If a particular sentence contains an 'i', it is reflective; otherwise, it is descriptive/non-reflective.

The marker-based approaches, outlined in the method Section 3.2.1 'Dictionary-based approaches', are essentially 'handcrafted' single rules (i.e., rules created manually). They work according to the rule: IF a word token of the word list occurs, THEN add one to the frequency count of the category represented by the word list.

Listing 2: OneR

```
IF      i:      < 0.5   THEN descriptive
IF      i:      >= 0.5  THEN reflective
```

The rule-based algorithms with higher reliability produced more complex rules than the OneR algorithm. For example, the oversampled C5.0 Rule model with the highest Cohens' κ created 44 rules. The listing below shows all the rules of this model. It contains several rules that determine whether a class is reflective (pos), but also rules that infer sentences as descriptive (see from rule 38 onwards). The default assumption for this rule set is that if none of the rules apply, the sentence is descriptive (neg). The rules are enumerated. The number in brackets indicates the number of instances covered by the rule. If the number of instances is followed by another number, the latter is the number of erroneously predicted instances by this rule.

For example, rule 2 covers 190 sentences of the training dataset. Of those, 189 sentences are reflective and one is descriptive. The rule reads as follows. If the unigram 'feel' is present, 'found' is absent, 'i' is present, 'includ' (includ is a stemmed unigram for the word include) is absent, and 'will' is absent, this sentence is reflective (pos).

Listing 3: C5.0 rules oversampled

```
Rule 1: (146)
felt > 0 AND i > 0 AND will <= 0 -> pos
Rule 2: (190/1)
feel > 0 AND found <= 0 AND i > 0 AND includ <= 0 AND will <= 0 -> pos
Rule 3: (82)
about > 0 AND explain <= 0 AND i > 0 AND it <= 0 AND see <= 0 AND think <= 0 AND
    will <= 0 -> pos
Rule 4: (77)
howev > 0 AND i > 0 AND question <= 0 AND will <= 0 -> pos
Rule 5: (60)
also <= 0 AND but <= 0 AND i > 0 AND team > 0 -> pos
```

```
Rule 6: (57)
i > 0 AND much > 0 -> pos
Rule 7: (34)
has > 0 AND my > 0 -> pos
Rule 8: (134/3)
me > 0 AND that > 0 -> pos
Rule 9: (32)
although > 0 AND i > 0 AND that <= 0 AND will <= 0 -> pos
Rule 10: (31)
and <= 0 AND i > 0 AND way > 0 -> pos
Rule 11: (64/1)
found > 0 AND i > 0 AND will <= 0 -> pos
Rule 12: (31)
i > 0 AND sure > 0 AND that <= 0 -> pos
Rule 13: (93/2)
i > 0 AND situat > 0 AND will <= 0 -> pos
Rule 14: (154/5)
i > 0 AND more > 0 AND will <= 0 -> pos
Rule 15: (50/1)
i > 0 AND peopl > 0 AND will <= 0 -> pos
Rule 16: (91/3)
i > 0 AND think > 0 AND will <= 0 -> pos
Rule 17: (88/3)
about <= 0 AND could > 0 AND i > 0 AND in <= 0 -> pos
Rule 18: (284/12)
i > 0 AND it > 0 AND more <= 0 AND think <= 0 AND which <= 0 AND will <= 0 -> pos
Rule 19: (129/5)
a <= 0 AND my > 0 AND that > 0 -> pos
Rule 20: (609/29)
i > 0 AND that > 0 AND will <= 0 -> pos
Rule 21: (570/28)
at <= 0 AND discuss <= 0 AND i > 0 AND that > 0 -> pos
Rule 22: (17)
but <= 0 AND never > 0 -> pos
Rule 23: (16)
after > 0 AND and <= 0 AND my > 0 -> pos
Rule 24: (16)
feel <= 0 AND i > 0 AND of > 0 AND practic > 0 -> pos
Rule 25: (15)
been > 0 AND may > 0 AND my <= 0 -> pos
Rule 26: (14)
been > 0 AND greater > 0 -> pos
Rule 27: (13)
and <= 0 AND i > 0 AND one > 0 AND that <= 0 AND was > 0 -> pos
Rule 28: (12)
been > 0 AND experi > 0 -> pos
Rule 29: (196/15)
me > 0 AND more <= 0 AND they <= 0 -> pos
Rule 30: (33/2)
been > 0 AND befor <= 0 AND i > 0 AND the <= 0 AND will <= 0 -> pos
Rule 31: (9)
i <= 0 AND like > 0 AND us > 0 -> pos
Rule 32: (19/1)
i > 0 AND idea > 0 AND that <= 0 AND think <= 0 -> pos
Rule 33: (42/4)
find > 0 AND i > 0 -> pos
Rule 34: (13/1)
fail > 0 AND my > 0 -> pos
Rule 35: (5)
avoid > 0 AND been > 0 -> pos
```

```
Rule 36: (5)
approach > 0 AND compani > 0 -> pos
Rule 37: (184/60)
but > 0 -> pos
Rule 38: (101/2)
a > 0 AND i <= 0 AND me <= 0 AND that > 0 AND us <= 0 -> neg
Rule 39: (22)
felt <= 0 AND found <= 0 AND take > 0 AND that <= 0 AND will <= 0 -> neg
Rule 40: (1015/46)
amount <= 0 AND approach <= 0 AND been <= 0 AND convers <= 0 AND i <= 0 AND me <=
    0 AND my <= 0 AND opinion <= 0 AND us <= 0 -> neg
Rule 41: (36/1)
avoid <= 0 AND been > 0 AND experi <= 0 AND greater <= 0 AND group <= 0 AND i <= 0
     AND may <= 0 -> neg
Rule 42: (9)
discuss > 0 AND i > 0 AND will > 0 -> neg
Rule 43: (30/2)
feel <= 0 AND felt <= 0 AND idea <= 0 AND more <= 0 AND question > 0 -> neg
Rule 44: (11/1)
at > 0 AND will > 0 -> neg
Rule 45: (1684/634)
that <= 0 AND think <= 0 -> neg
```

6.1.3 *Results of the high performance models*

As with the models previously described, the high performing models are trained and tested with the dataset described in Section 6.1, and according to the process outlined in Section 4.6.2 'Research design'. The description of the models can be found in Section 4.5.3 'High performance models', and the metrics are described in Section 4.2 'Evaluation criteria and metrics'.

Table 17 shows the reliability achieved for these models. Three of the models have a Cohen's κ of 0.7. They are the SVM with polynomial kernel, and both Random Forest models. The four models SVM linear, SVM radial, oversampled SVM radial, and oversampled Naïve Bayes have a Cohen's κ of 0.69. All models have a Gwet's AC_1 from 0.8 to 0.84. The per cent agreement ranges from 87% to 89%, which means that the models disagree with the labels of the test dataset in 11% to 13% of the cases.

Method	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
SVM linear	456	0.89	0.69	0.69	0.82	0.69
SVM linear oversampled	456	0.87	0.64	0.64	0.80	0.64
SVM radial	456	0.89	0.69	0.69	0.82	0.69
SVM radial oversampled	456	0.89	0.69	0.68	0.82	0.69
SVM polynomial	456	0.89	0.70	0.70	0.83	0.70
SVM polynomial oversampled	456	0.88	0.63	0.63	0.81	0.63
NNET	456	0.89	0.68	0.67	0.82	0.67
NNET oversampled	456	0.88	0.67	0.67	0.80	0.67
Random Forest	456	0.89	0.70	0.70	0.84	0.70
Random Forest oversampled	456	0.89	0.70	0.70	0.81	0.70
Naïve Bayes	456	0.88	0.66	0.66	0.80	0.66
Naïve Bayes oversampled	456	0.88	0.69	0.69	0.81	0.69

Table 17: Reliability of high performance models

Table 18 shows the performance measures AUC, sensitivity (Sens.) and specificity (Spec.). All models but one have a higher AUC than 0.9. The oversampled Random Forest has the highest sensitivity, whereas SVM with polynomial kernel function on the oversample dataset achieves the highest specificity. The Random Forest (oversampled) has approximately the same sensitivity and specificity as the Naïve Bayes (oversampled).

Method	N	AUC	Sens.	Spec.
SVM linear	456	0.93	0.71	0.95
SVM linear oversampled	456	0.89	0.66	0.94
SVM radial	456	0.93	0.71	0.95
SVM radial oversampled	456	0.91	0.70	0.95
SVM polynomial	456	0.93	0.73	0.95
SVM polynomial oversampled	456	0.93	0.60	0.97
NNET	456	0.93	0.66	0.96
NNET oversampled	456	0.91	0.73	0.93
Random Forest	456	0.94	0.69	0.96
Random Forest oversampled	456	0.94	0.79	0.92
Naïve Bayes	456	0.93	0.72	0.93
Naïve Bayes oversampled	456	0.92	0.78	0.92

Table 18: Performance measures of high performance models

Figure 11 shows the ROC curves for the SVM polynomial, Random Forest, and Naïve Bayes.

AUC is high for all three models. Their ROC curves are extremely similar. Again, the trade-off between sensitivity and specificity can be observed. The dots on the ROC curve represent the probability cut-off that would likely yield the highest sum of sensitivity and specificity if trained and tested on new data. We can see that, by changing the probability threshold of the models, specificity can be reduced in order to gain higher sensitivity, which is over 0.85 for these thresholds. Furthermore, the curves reach nearly the top of the 1.0 sensitivity threshold, indicating that extremely high sensitivity would be within reach by reducing specificity.

6.1.4 *Discussion of the results of the three lines of investigation*

Section 6.1 'Reflection' reported the results of three lines of investigation into the problem of detecting sentences that are reflective and those that are descriptive. In each line of investigation, several machine learning algorithms are trained and tested on the same dataset of reflective and descriptive sentences.

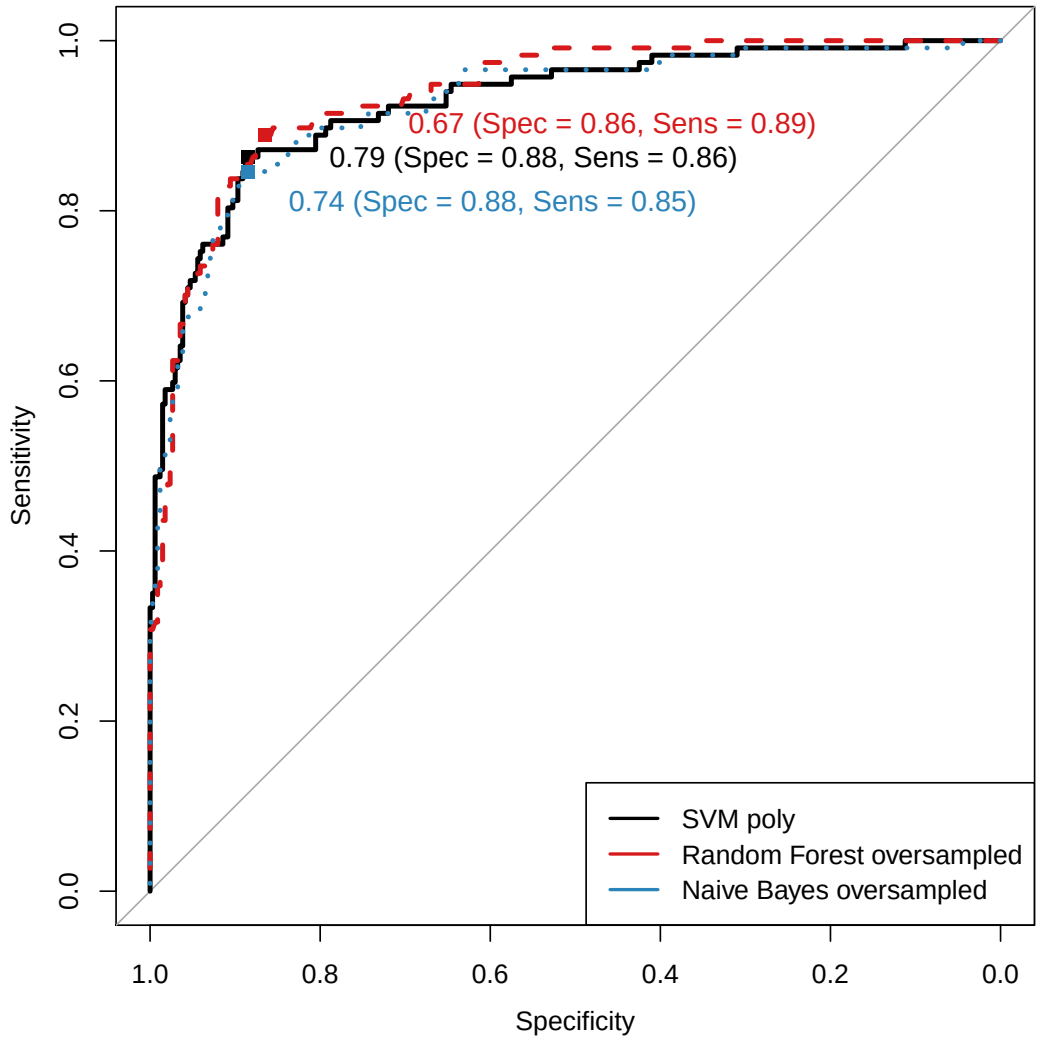

Figure 11: ROC-curves of high performance models

For each line of investigation, the models are assessed with regard to their reliability and performance. In addition, the ROC curve for the three models of each line is inspected, and alternative probability cut-off points are discussed.

Section 6.1.1 'Results of the tree-based models' describes the results of the first line of investigation. In total, 12 models from four different machine learning algorithms are assessed. In addition, the tree of one model is visualised and inspected.

Section 6.1.2 'Results of the rule-based models' presents the results of the second line of investigation. The models of the four rule-based algorithms are evaluated. In addition, the rule sets of two models are listed.

Section 6.1.3 'Results of the high performance models' outline the reliability and models' performance measures generated by the machine learning algorithms SVM, Neural Network, Random Forest, and Naïve Bayes.

In the last three sections, the tree-based, rule-based, and several high performance machine learning algorithms are evaluated with regard to their potential for detecting reflection. All models are evaluated under the same conditions outlined in Section 4.6.2 'Research design'. This allows comparing the models of the three lines of investigation with regard to their potential for detecting reflection.

Table 19 shows the top three models of each line of investigation. They are ranked by Cohen's κ. In addition to Cohen's κ, the table shows the corresponding per cent agreement, sensitivity, and specificity.

Line of investigation	Method	Cohen's κ	%	Sensitivity	Specificity
High performance models	Random Forest	0.70	0.89	0.69	0.96
	Random Forest oversampled	0.70	0.89	0.79	0.92
	SVM polynomial	0.70	0.89	0.73	0.95
Rule-based models	C5.0 Rules oversampled	0.65	0.87	0.74	0.91
	JRip	0.63	0.86	0.68	0.92
	C5.0 Rules	0.59	0.84	0.70	0.89
Tree-based models	CART Tree cost	0.64	0.87	0.72	0.92
	CART Tree maxdepth	0.64	0.87	0.72	0.92
	Conditional Inference Tree mincriterion	0.64	0.87	0.72	0.92

Table 19: Top models of each line of investigation for the indicator reflection

The top performers of the high performance models have all higher Cohen's κ compared to the rule-based and tree-based top performers. The rule-based and tree-based models are on a similar level.

From all investigated rule-based models, the C5.0 Rules algorithm trained on the oversampled data has the highest Cohen's κ at 0.65, followed by JRip with 0.63, and the C5.0 Rules model trained on the training data with the original distribution. The top C5.0 Rules model also has the highest sensitivity and the second highest specificity compared with the other two models.

The findings show that most models have lower sensitivity than specificity. Proportionally more sentences that are reflective are marked as descriptive. Inspection of the ROC curves indicates that, to a degree, sensitivity can be improved with an acceptable loss in specificity. The ROC curve of the tree-based and rule-based models degraded more quickly when defining alternative probability cut-off thresholds in order to trade specificity for higher sensitivity. The curves of the high performing models indicate that, to a degree, higher sensitivity can be achieved without a disproportional large decrease in specificity.

According to the benchmarks reported in Section 4.2 'Evaluation criteria and metrics', the Cohen's κ reliability values of the tree-based, rule-based, and high performance models are all substantial (Landis and Koch, 1977). Most high performing models have a Krippendorff's α that allows for tentative conclusions (Krippendorff, 2012). The tree-based and rule-based models do not reach this threshold. However, none of the models has almost perfect κ values. According to the reliability values condensed from the research on the content analysis of reflective writing (see Section 3.1.4 'Manual reflection detection performance'), the per cent agreement of most models for the three lines of investigation are in the middle bracket, which is between 80% and 90%. The top performing high performance models came close to the top bracket starting at 90%. All models are above the 0.5

Cohen's κ threshold for exploratory research (Stemler and Tsai, 2008), and the benchmark of Fleiss et al. (2003) places all models in the range from fair to good. Overall, the benchmarks indicate that reflection can be detected reliably. In particular, the Random Forest and SVM polynomial achieve high reliability benchmark values, which is already suitable for many scenarios. However, they do not achieve the highest benchmark values, which excludes their applicability for situations with strict requirements on reliability.

The dataset has an estimated per cent agreement of 97% and a Cohen's κ of 0.92 (see Table 9 in Section 5.7.4 'Reliability'). Compared with the standard set by raters, the best high performing machine learning algorithm has 8% lower agreement and 0.22 points lower Cohen's κ. According to the benchmark of Landis and Koch (1977), the almost perfect reliability of the raters drops to substantial for the models. This lower reliability of the machine learning models compared with the human standard is in line with the research outlined in Section 3.2.3 'Machine learning approaches'. As outlined there, McKlin (2004) reported for the dataset a Cohen's κ of 0.85 and a κ of 0.7 for the model with the highest performance. The model of Corich (2011) achieved a κ of 0.68 based on a dataset with a κ of 0.82. The decrease in reliability reported there is smaller than for the best model reported here. Both papers described several optimisation strategies in order to maximise model performance.

The tree-based and rule-based models have a similar range of κ values. However, the rule-based models range slightly higher than the tree-based models. The Cohen's κ values for the tree-based models range from 0.56 to 0.64, whereas the rule-based models range from 0.58 to 0.65[2]. Such slightly better performance of the rule-based models is in line with expectations for the performance of these models. The rule-based approach is more flexible when modelling the data because it does not rely on a single root node to make all subsequent classification decisions. All the high

2 OneR is excluded because it serves only for demonstrating purposes (see the description for OneR in Section 4.5.2 'Rule-based models').

performance models use different strategies to build their models from the data. For example, the Random Forest algorithm, which produced a model in the top range of this line of investigation, generates many tree models based on a random subset of the data. Each of the generated trees models the data in slightly different ways. The final decision on the label is based on an aggregation of the tree classifications. This strategy is more flexible than modelling the data with only one tree, and it does improve model performance.

One of the benefits of the tree-based and rule-based models is that these models are more accessible to interpretation because they follow either the notion of decision trees or rule sets. An observation that can be made from the inspection of the tree-based and two rule-based models is that simple models already achieve reasonable results, but in order to reach higher reliabilities, model complexity increases quickly. For example, inspection of the model for the OneR algorithm, which creates an extremely simple model that classifies sentences based only on the presence or absence of the unigram 'i', has a Cohen's κ of 0.51. This is in line with the research of (Birney, 2012, p. 269) that found a strong relationship between use of the first person voice and reflection with a linguistic study. The tree-based model of the Conditional Inference Tree considers more decisions in order to achieve a Cohen's κ of 0.64. The rule-based model of the C5.0 Rule algorithm creates 44 rules in order to achieve a κ of 0.65. The strong performing machine learning algorithms reaches a κ of 0.7, which is 0.05 points higher than the C5.0 Rule model. This suggests that simple strategies can already achieve moderate classification results. The entry level to automatically detect reflection is relatively low. However, higher levels of reliability are achieved only with the high performing algorithms.

The inspected tree-based and rule-based models show that their set of model features is relatively small. The tree-based model bases the decision to classify sentences on seven features, whereas the rule-based model uses 60 features. This relatively small

set of information bearing features achieves already usable classifications. This is in line with the research outlined in Section 3.2.1 'Dictionary-based approaches'. The approach described there is based on dictionaries with manually constructed word lists that represent a category. This research shows that dictionaries with carefully selected words can be used to detect patterns from text.

Unlike the dictionary-based approaches, the tree-based and rule-based models form decision trees or rules with these features. Their working is similar to the research outlined in Section 3.2.2 'Section 3.2.2' that used handcrafted rules in order to detect meaningful patterns in text. Compared with the dictionary-based approach, rule-based approaches allow associating words with rules, which has the benefit of modelling the data more precisely. For example, Ullmann et al. (2012) constructed several rules that approximated reflection. Similarly to the inspected rule set, each rule had several conditions. However, these rules were handcrafted by the researcher according to the theoretical considerations of what constitutes reflection, and not automatically learned from data.

It is notable that the rule set contains not only many self-references, such as 'I', 'me', and 'my', but also references to feelings and thinking. There are also links to words that contrast a statement indicating a critical stance, for example, 'but', 'however', and 'although'. Included are also words that de-emphasise absolute statements, such as 'could' and 'may', and words that refer to the writer experience and context, for example, 'situation', 'experience', 'practice', 'approach', and 'fail'. Although these rules are generated automatically from data, they entail some of the essence of written reflection.

6.2 COMMON CATEGORIES OF REFLECTIVE WRITING

This section presents the evaluation results of the machine learning algorithms that answer the second research question, **Q2: Can machine learning algorithms be used to detect common categories of reflective writing?** Similarly to Section 6.1 'Reflection', evaluation of the machine learning algorithms is performed under the standardised conditions described in the research design (see Section 4.6.2 'Research design', Figure 3, and Figure 4). In the previous section, all machine learning models are trained and tested on the same dataset. In this section, each indicator of the common categories of reflection has its own set of data. The datasets are the result of the data generation process described in Section 4.6.1 'Dataset generation process'. Their implementation is described in Chapter 5 'DATASET GENERATION', and an overview of all datasets can be found in Section 5.7.6 'Quality standard and datasets statistics'. Examples of all the indicators of the common categories of reflective writing are prepared in Appendix F 'EXAMPLES OF THE DATASETS'. These datasets are used to train the models and evaluate their reliability on a test dataset. The machine learning algorithms chosen in order to evaluate their performance on the task for detecting common categories of reflection are described in Section 4.5.3 'High performance models'. These algorithms are used for the third line of investigation outlined in Section 6.1.3 'Results of the high performance models'. The results there suggest that the high performance models are good candidate models for the problem to detect reflection (see Section 6.1.4 'Discussion of the results of the three lines of investigation'). The focus results described here shifts from evaluation of the differences in the types of machine learning algorithms (tree-based, rule-based, and

high performing models) to a comparison of the differences in the same machine learning algorithms on different datasets.

The structure of the following sections follows the order of the common categories described in Section 2.3.2 'Common reflection categories'. Most categories are measured with one indicator. The results of the two indicators for the category **Outcome** are reported in the same section (see the mapping of indicators to categories in Section 5.7.2 'Task design'). Discussion on the findings can be found in Section 6.2.7 'Discussion of the results of the common categories of reflection'.

6.2.1 *Description of an experience*

Table 20 shows the details of the training dataset. The amount of positive instances (presence of **Experience**) is close to the amount of negative instances (absence of **Experience**). The dataset is more balanced than for the indicator **Reflection**. The oversampled dataset is generated by randomly duplicating positive instances until the amount of positive instances matches the amount of negative instances (see Section 4.6.2 'Research design'). Some examples of the dataset for the indicator **Experience** can be found in Table 40 in Appendix F 'EXAMPLES OF THE DATASETS'.

Statistics	Count
Number of features	950
Number of instances	3272
Number of training instances	2618
Number of positive training instances	1204
Number of negative training instances	1414
Number of test instances	654
Number of positive test instances	301
Number of negative test instances	353

Table 20: Statistics about the training and test set of the indicator Experience

Table 21 shows the reliability of the machine learning models. Reliability is determined from the test dataset. The SVM linear model, SVM radial model (both with the original instance distribution and oversampled instance distribution), and SVM with polynomial kernel function have the highest κ at 0.83. The lowest Cohen's κ is 0.79 for both SVM with linear kernel trained on the oversampled data and the Naïve Bayes classifier. For all models, the agreement is over 90%. All other reliability measures are extremely close to the value of Cohen's κ. It is notable that reliability values for Gwet's AC_1 are close to the other reliability measures compared with the results reported in Section 6.1.3 'Results of the high performance models'. This can be expected because Gwet (2008) reported that, for balanced data, the AC_1 values are similar to the other reliability measures.

Method	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
SVM linear	654	0.92	0.83	0.83	0.83	0.83
SVM linear upsampled	654	0.89	0.79	0.79	0.79	0.79
SVM radial	654	0.91	0.83	0.83	0.83	0.83
SVM radial upsampled	654	0.92	0.83	0.83	0.84	0.83
SVM polynominal	654	0.91	0.83	0.83	0.83	0.83
SVM polynominal upsampled	654	0.90	0.80	0.80	0.81	0.80
NNET	654	0.91	0.82	0.82	0.82	0.82
NNET upsampled	654	0.90	0.81	0.81	0.81	0.81
Random Forest	654	0.91	0.82	0.82	0.83	0.82
Random Forest upsampled	654	0.90	0.81	0.81	0.81	0.81
Naïve Bayes	654	0.90	0.79	0.79	0.79	0.79
Naïve Bayes upsampled	654	0.90	0.80	0.80	0.80	0.80

Table 21: Reliability of indicator Experience

6.2.2 *Feelings*

Table 22 shows descriptive statistics for the dataset of the indicator **Feelings**. From the distribution of the positive and negative instances, it can be seen that the dataset is imbalanced. There are more instances that indicate absence (negative instances) than presence of **Feelings** (positive instances). Table 41 in Appendix F 'EXAMPLES OF THE DATASETS' lists examples for both classes.

Statistics	Count
Number of features	811
Number of instances	2606
Number of training instances	2085
Number of positive training instances	629
Number of negative training instances	1456
Number of test instances	521
Number of positive test instances	157
Number of negative test instances	364

Table 22: Statistics about the training and test set of the indicator Feelings

Table 23 reports the reliability measures of the machine learning algorithms. The oversampled Random Forest has a Cohen's κ of 0.73 followed by the Random Forest trained on training data with the original class distribution (κ = 0.72). The linear SVM on the oversampled data has the lowest κ at 0.64. Agreement is generally high, and ranges from 85% to 89%.

Method	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
SVM linear	521	0.87	0.69	0.69	0.78	0.69
SVM linear oversampled	521	0.85	0.64	0.64	0.74	0.64
SVM radial	521	0.87	0.69	0.69	0.78	0.69
SVM radial oversampled	521	0.87	0.70	0.70	0.78	0.70
SVM polynominal	521	0.87	0.68	0.68	0.78	0.68
SVM polynominal oversampled	521	0.86	0.65	0.65	0.77	0.65

Continued ...

Method	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
NNET	521	0.87	0.69	0.69	0.78	0.69
NNET oversampled	521	0.85	0.65	0.65	0.74	0.65
Random Forest	521	0.89	0.72	0.72	0.81	0.72
Random Forest oversampled	521	0.88	0.73	0.73	0.80	0.73
Naïve Bayes	521	0.86	0.67	0.67	0.76	0.67
Naïve Bayes oversampled	521	0.87	0.69	0.69	0.77	0.69

Table 23: Reliability of indicator Feelings

6.2.3 *Personal*

For the common category **Personal**, the indicator **Beliefs** is developed (see Section 5.7.2 'Task design'). In the datasets discussed thus far, the instances of interest are in the minority. For this dataset, more instances are positive than negative. Therefore, the oversampled dataset randomly duplicates negative instances in order to balance the class distribution. Table 42 in Appendix F 'EXAMPLES OF THE DATASETS' shows examples of these instances.

Statistics	Count
Number of features	709
Number of instances	2253
Number of training instances	1804
Number of positive training instances	928
Number of negative training instances	876
Number of test instances	449
Number of positive test instances	231
Number of negative test instances	218

Table 24: Statistics about the training and test set of the indicator Beliefs

Table 25 shows the reliability for the machine learning algorithms. SVM linear achieves the highest Cohen's κ at 0.66, followed by SVM radial trained on the oversampled data and NNET, both with a κ of 0.65. All reported Cohen's κ values are in the range of 0.61 to 0.66. Their difference in the per cent agreement is relatively small: it ranges from 81% to 83%.

Method	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
SVM linear	449	0.83	0.66	0.66	0.66	0.66
SVM linear oversampled	449	0.81	0.61	0.61	0.61	0.61
SVM radial	449	0.82	0.64	0.64	0.64	0.64
SVM radial oversampled	449	0.82	0.65	0.65	0.65	0.65
SVM polynominal	449	0.82	0.64	0.64	0.64	0.64
SVM polynominal oversampled	449	0.82	0.63	0.63	0.63	0.63
NNET	449	0.82	0.65	0.65	0.65	0.65
NNET oversampled	449	0.82	0.64	0.64	0.64	0.64
Random Forest	449	0.82	0.63	0.63	0.64	0.63
Random Forest oversampled	449	0.82	0.64	0.64	0.64	0.64
Naïve Bayes	449	0.81	0.63	0.63	0.63	0.63
Naïve Bayes oversampled	449	0.81	0.62	0.62	0.62	0.62

Table 25: Reliability of indicator Beliefs

6.2.4 *Critical stance*

One indicator is developed for this common category of reflective writing, namely the indicator **Difficulties** (see Section 5.7.2 'Task design'). Table 43 in Appendix F 'EXAMPLES OF THE DATASETS' lists several examples of this dataset.

Table 26 shows statistics that describe the dataset for the indicator **Difficulties**. From the distribution of positive and negative, it can be seen that the dataset has a balanced class distribution.

Statistics	Count
Number of features	808
Number of instances	2630
Number of training instances	2104
Number of positive training instances	1072
Number of negative training instances	1032
Number of test instances	526
Number of positive test instances	268
Number of negative test instances	258

Table 26: Statistics about the training and test set of the indicator Difficulties

Table 27 reports the reliability of all machine learning models. Both SVM with radial kernel function and polynomial SVM have the highest Cohen's κ at 0.60. All machine learning algorithms are in the range of 0.55 to 0.60. The per cent agreement is between 77% and 80%.

Method	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
SVM linear	526	0.77	0.55	0.55	0.55	0.55
SVM linear oversampled	526	0.79	0.59	0.59	0.59	0.59
SVM radial	526	0.80	0.60	0.60	0.60	0.60
SVM radial oversampled	526	0.78	0.56	0.56	0.56	0.56
SVM polynominal	526	0.80	0.60	0.60	0.60	0.60
SVM polynominal oversampled	526	0.80	0.59	0.59	0.59	0.59
NNET	526	0.79	0.58	0.58	0.58	0.58
NNET oversampled	526	0.79	0.58	0.58	0.58	0.58
Random Forest	526	0.79	0.58	0.58	0.58	0.58
Random Forest oversampled	526	0.79	0.57	0.57	0.57	0.57
Naïve Bayes	526	0.79	0.59	0.59	0.59	0.59
Naïve Bayes oversampled	526	0.79	0.57	0.57	0.57	0.58

Table 27: Reliability of indicator Difficulties

6.2.5 *Perspective*

For this common category of reflection, one indicator is tested. Table 28 shows the amount of instances for training and testing produced by the data generation process. From the table, we can see that class distribution is unbalanced. Examples for the dataset of indicator **Perspective** can be found in Table 44 in Appendix F 'EXAMPLES OF THE DATASETS'.

Statistics	Count
Number of features	591
Number of instances	1983
Number of training instances	1587
Number of positive training instances	246
Number of negative training instances	1341
Number of test instances	396
Number of positive test instances	61
Number of negative test instances	335

Table 28: Statistics about the training and test set of the indicator Perspective

Table 29 shows reliability for all the models. The highest Cohen's κ of 0.55 was achieved by the Naïve Bayes model, followed with a κ of 0.52 by the SVM polynomial. All other κ values are below 0.5, with the lowest κ at 0.3. Whereas the κ values are relatively low, the agreement is relatively high. All models agree in 83% of the cases or more. The highest per cent agreement of 89% is achieved by the SVM polynomial model. The AC_1 values range from 0.77 to 0.84.

Method	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
SVM linear	396	0.88	0.46	0.45	0.84	0.46
SVM linear oversampled	396	0.83	0.32	0.32	0.77	0.32
SVM radial	396	0.88	0.46	0.45	0.84	0.46
SVM radial oversampled	396	0.84	0.37	0.37	0.79	0.37

Continued ...

Method	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
SVM polynominal	396	0.89	0.52	0.52	0.85	0.52
SVM polynominal oversampled	396	0.87	0.29	0.26	0.84	0.29
NNET	396	0.87	0.44	0.43	0.83	0.44
NNET oversampled	396	0.85	0.40	0.40	0.81	0.40
Random Forest	396	0.87	0.30	0.28	0.84	0.30
Random Forest oversampled	396	0.86	0.30	0.28	0.82	0.30
Naïve Bayes	396	0.88	0.55	0.55	0.84	0.55
Naïve Bayes oversampled	396	0.85	0.48	0.47	0.78	0.48

Table 29: Reliability of indicator Perspective

6.2.6 *Outcome*

To inspect the **Outcome** common category of reflective writing, two indicators are developed, namely the indicators **Intention** and **Learning**. First, we discuss the results for the indicator **Intention**, and then the results for the indicator **Learning**.

Table 30 shows the descriptive statistics of the dataset for the indicator **Intention**. The class distribution is unbalanced. A total of 2,639 training instances that do not express **Intention** are compared with 273 instances that express **Intention**.

Table 45 in Appendix F 'EXAMPLES OF THE DATASETS' shows several examples of the dataset.

Statistics	Count
Number of features	1028
Number of instances	3636
Number of training instances	2909
Number of positive training instances	273
Number of negative training instances	2636
Number of test instances	727
Number of positive test instances	68
Number of negative test instances	659

Table 30: Statistics about the training and test set of the indicator Intention

Table 31 reports reliabilities. Cohen's κ ranges from 0.49 to 0.71. Both the Naïve Bayes and the Random Forest model have the highest κ of 0.71, followed by the Random Forest trained on the oversampled data with a κ of 0.70. The per cent agreement ranges from 93% to 95%.

Method	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
SVM linear	727	0.95	0.69	0.69	0.94	0.69
SVM linear oversampled	727	0.93	0.55	0.55	0.92	0.55
SVM radial	727	0.95	0.66	0.66	0.94	0.66
SVM radial oversampled	727	0.94	0.62	0.62	0.93	0.62
SVM polynominal	727	0.95	0.66	0.66	0.94	0.66
SVM polynominal oversampled	727	0.94	0.49	0.48	0.93	0.49
NNET	727	0.95	0.64	0.64	0.94	0.64
NNET oversampled	727	0.94	0.65	0.65	0.92	0.65
Random Forest	727	0.95	0.71	0.71	0.94	0.71
Random Forest oversampled	727	0.95	0.70	0.70	0.94	0.71
Naïve Bayes	727	0.95	0.71	0.71	0.95	0.71
Naïve Bayes oversampled	727	0.93	0.65	0.65	0.92	0.66

Table 31: Reliability of indicator Intention

The dataset for the indicator **Learning** is compared with the dataset for the indicator **Learning** less unbalanced. Table 32 shows that, altogether, 544 positive

training instances and 920 negative training instances are available. Examples of these instances can be found in Table 46 in Appendix F 'EXAMPLES OF THE DATASETS'.

Statistics	Count
Number of features	595
Number of instances	1828
Number of training instances	1464
Number of positive training instances	544
Number of negative training instances	920
Number of test instances	364
Number of positive test instances	135
Number of negative test instances	229

Table 32: Statistics about the training and test set of the indicator Learning

Table 33 shows the reliabilities of the machine learning algorithms on the test dataset for the indicator **Learning**. Both SVM linear and SVM polynomial have the highest Cohen's κ at 0.63, followed by the SVM radial model (0.62). The SVM linear trained on the oversampled data has the lowest κ at 0.49. The per cent agreement ranges from 77% to 83%.

Method	N	%	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)
SVM linear	364	0.83	0.63	0.63	0.69	0.63
SVM linear oversampled	364	0.77	0.49	0.49	0.57	0.49
SVM radial	364	0.83	0.62	0.62	0.68	0.62
SVM radial oversampled	364	0.80	0.56	0.56	0.62	0.56
SVM polynominal	364	0.83	0.63	0.63	0.69	0.63
SVM polynominal oversampled	364	0.79	0.55	0.55	0.62	0.55
NNET	364	0.80	0.57	0.58	0.64	0.58
NNET oversampled	364	0.77	0.50	0.50	0.58	0.50
Random Forest	364	0.82	0.60	0.60	0.68	0.60
Random Forest oversampled	364	0.82	0.60	0.60	0.66	0.60
Naïve Bayes	364	0.82	0.61	0.61	0.66	0.61
Naïve Bayes oversampled	364	0.80	0.58	0.58	0.64	0.58

Table 33: Reliability of indicator Learning

6.2.7 *Discussion of the results of the common categories of reflection*

Section 6.2 'Common categories of reflective writing' presents the assessment results of the models on the seven indicators for the six common categories of reflection. Compared with the previous section, where one dataset is evaluated with many different machine learning algorithms, here, many different datasets are evaluated with the same set of machine learning algorithms. For each indicator, the models for the SVM, Neural Network, Random Forest, and Naïve Bayes algorithms are trained and tested. For each dataset and model, several reliability indices are reported based on the test dataset.

Similarly to Section 6.1.4 'Discussion of the results of the three lines of investigation', this section compares the machine learning algorithms across all common reflection categories.

For all investigated indicators, Table 34 shows the top three models ranked by Cohen's κ. Several **general observations** can be made.

First, it can be seen that within each indicator group, the Cohen's κ values of the top models are close to each other. This suggests that the reliabilities are relatively stable for most indicators. Their κ values vary by 0.03 at the most, with one exception: the models for the indicator **Perspective**. There, the maximum difference is 0.07 between the third model and the first two models. Overall, the κ values of the top three models for each indicator dataset are stable.

Second, Table 34 also shows that there is not a single best performing machine learning algorithm across all indicators. More frequently, all three versions of the SVM and Random Forest are in the top position. Less frequent, but important, is the Naïve Bayes because its models reach the top position for the indicators **Intention** and **Perspective**.

Third, the chosen data-driven strategy to counter the class imbalance problem with random oversampling does improve the reliability of the machine learning algorithms in many cases. In all cases but one, at least one model trained with oversampled data is present in the top three lists.

Indicator	Method	Cohen's κ	Krippendorff's α	Gwet's AC_1	ICC(1,1)	%
Experience	SVM radial oversampled	0.83	0.83	0.84	0.83	0.92
	SVM linear	0.83	0.83	0.83	0.83	0.92
	SVM polynominal	0.83	0.83	0.83	0.83	0.91
Feelings	Random Forest oversampled	0.73	0.73	0.80	0.73	0.88
	Random Forest	0.72	0.72	0.81	0.72	0.89
	SVM radial oversampled	0.70	0.70	0.78	0.70	0.87
Beliefs	SVM linear	0.66	0.66	0.66	0.66	0.83
	SVM radial oversampled	0.65	0.65	0.65	0.65	0.82
	NNET	0.65	0.65	0.65	0.65	0.82
Difficulties	SVM radial	0.60	0.60	0.60	0.60	0.80
	SVM polynominal	0.60	0.60	0.60	0.60	0.80
	SVM polynominal oversampled	0.59	0.59	0.59	0.59	0.80
Perspective	Naïve Bayes	0.55	0.55	0.84	0.55	0.88
	SVM polynominal	0.52	0.52	0.85	0.52	0.89
	Naïve Bayes oversampled	0.48	0.47	0.78	0.47	0.85
Intention	Naïve Bayes	0.71	0.71	0.95	0.71	0.95
	Random Forest	0.71	0.71	0.94	0.71	0.95
	Random Forest oversampled	0.70	0.70	0.94	0.70	0.95
Learning	SVM linear	0.63	0.63	0.69	0.63	0.83
	SVM polynominal	0.63	0.63	0.69	0.63	0.83
	SVM radial	0.62	0.62	0.68	0.62	0.83
Reflection	Random Forest	0.70	0.70	0.84	0.70	0.89
	Random Forest oversampled	0.70	0.70	0.81	0.70	0.89
	SVM polynomial	0.70	0.70	0.83	0.70	0.89

Table 34: Top models of all indicators of reflection

Section 4.2 'Evaluation criteria and metrics' outlines several **benchmarks** for acceptable reliability thresholds. In addition to these generic benchmarks, Section 3.1.4 'Manual reflection detection performance' distils the reliability ranges

found in the research for the analysis of reflective writing (see page 69). For all top models, Table 35 shows their respective levels with regard to the benchmarks. The table shows the number of test instances (N), Cohen's κ, Krippendorff's α (both have the same values on the second digit), per cent agreement, model level according to the benchmark of Landis and Koch (1977), Krippendorff (2012) (tent. means 'allows tentative conclusions' and not rel. means 'not reliable'), and per cent agreement bracket (top: >= 90%, middle: 80-90%, low: 50-80%) derived from the research on the analysis of reflective writing.

Indicator	N	Cohen's κ & Kripp. α	%	Landis and Koch	Krippendorff	% Raters
Experience	654	0.83	0.92	almost perfect	reliable	top
Feelings	521	0.73	0.88	substantial	tent.	middle
Beliefs	449	0.66	0.83	substantial	not rel.	middle
Difficulties	526	0.60	0.80	moderate	not rel.	middle
Perspective	396	0.55	0.88	moderate	not rel.	middle
Intention	727	0.71	0.95	substantial	tent.	top
Learning	364	0.63	0.83	substantial	not rel.	middle
Reflection	456	0.70	0.89	substantial	tent.	middle

Table 35: Benchmarks of top models of all indicators

The indicator **Experience** has the highest benchmark level for all benchmarks. Four of the indicators are benchmarked as substantial, and two as moderate, on the benchmark of Landis and Koch (1977).

According to the benchmark of Krippendorff (2012), **Experience** is reliable, and both **Feelings** and **Intention** allow for tentative conclusions. Four of the indicators are not reliable.

Compared with the per cent agreements of the raters reported in Section 3.1.4 'Manual reflection detection performance', five of the indicator models are in the middle, and two are in the top bracket. None are in the low bracket, although the model for the indicator **Difficulties** is close to this threshold.

All indicator models are in the 'fair-to-good' range reported by Fleiss et al. (2003), with the exception of **Experience**, which is excellent.

The Cohen's κ values are all above the 0.5 threshold for exploratory research described by Stemler and Tsai (2008).

However, compared with the **reliability of the datasets**, model reliability is lower. Rater reliability for the dataset is estimated in Section 5.7.4 'Reliability' on a sample of the datasets, and summarised in Table 9.

Table 36 shows the Cohen's κ values of the top model for each indicator (Model), along with the κ of the datasets (Dataset) and the difference between both (Difference). The agreements for this model and the dataset are shown next to the values.

	Reliability			Agreement		
Indicator	Model	Dataset	Difference	Model	Dataset	Difference
Experience	0.83	0.96	0.13	0.92	0.98	0.06
Feelings	0.73	0.94	0.21	0.88	0.98	0.10
Beliefs	0.66	0.93	0.27	0.83	0.97	0.14
Difficulties	0.60	0.94	0.34	0.80	0.97	0.17
Perspective	0.55	0.78	0.23	0.88	0.95	0.07
Intention	0.71	0.93	0.22	0.95	0.99	0.04
Learning	0.63	0.91	0.28	0.83	0.96	0.13
Reflection	0.70	0.92	0.22	0.89	0.97	0.08

Table 36: Reliability of models and datasets

From Table 36, we can see that dataset reliability for all indicators but one is above 0.90. Model reliability is generally lower than dataset reliability. On average, the

difference is 0.24. The indicator **Difficulties** shows the largest decrease in reliability, whereas the indicator **Experience** has the smallest drop.

On average, model agreement is 10% lower than that of the raters. The smallest difference is for the indicator **Intention**, and the largest for **Difficulties**.

In general, Table 36 shows that the machine learning algorithms have lower reliability values than the datasets rated by the four-fifth majority vote rater. As described in Section 6.1.4 Discussion of the results of the three lines of investigation, this general lower model reliability compared with the dataset is also found in the literature described in Section 3.2.3 'Machine learning approaches'.

This drop in the reliability measures can be expected, and has consequences for the general applicability of machine learning for the models of reflective writing outlined in Section 2.2 'Models to analyse written reflection'. The indicator datasets have reliability and agreement in the top brackets of the measures summarised in Section 3.1.4 'Manual reflection detection performance'. Several of the papers reported there have an agreement in the top bracket, from 90% and above. The machine learning models created on the categories reported there would have a lower agreement. The extent of the decrease will vary, but can be expected.

The extent to which the reliability of the machine learning models is lower compared with the dataset also depends on algorithm optimisation. As outlined in Section 4.6.2 'Research design', the approach taken here is to train and test the machine learning methods under standardised conditions in order to inspect their differences. This also means that the models were not individually optimised. For example, McKlin (2004) and Corich (2011) reported Cohen's κ values that were 0.15 and 0.14 points below dataset reliability. McKlin (2004) described several cycles of inspecting the models and the use of a dictionary in order to achieve these reliabilities.

6.3 SUMMARY

This chapter implemented the research design outlined in Section 4.6.2 'Research design'. It reported the evaluation results, and discussed its findings. The sections were structured according to the two research questions posed in Section 1.1 'Research questions'. Section 6.1 'Reflection' reported and discussed the results for each of the three lines of investigations used to structure the argument for the first research question, and Section 6.2 'Common categories of reflective writing' presented and discussed the results relevant to the second research question.

The results of both sections were based on the datasets generated for each reflection indicator. Each dataset was generated with the same data generation process outlined in Section 4.6.1 'Dataset generation process'. Whereas for the Section 6.1 'Reflection', all three types of machine learning algorithms were evaluated on the same dataset, Section 6.2 'Common categories of reflective writing' assessed the models of the high performing line of investigation on seven different datasets. Each dataset contained instances of the reflection indicators. The model training and assessment for all machine learning algorithms was conducted under the same conditions outlined in Section 4.6.2 'Research design'.

The main findings were discussed in Section 6.1.4 'Discussion of the results of the three lines of investigation' and Section 6.2.7 'Discussion of the results of the common categories of reflection', and are summarised here and brought into the context of the data generation process findings.

Overall, most models were able to reliably detect reflection. According to the benchmark of Landis and Koch (1977), most models had a substantial or higher reliability. Compared with the reported per cent agreement from the research of the

analysis of reflective writing, all models were in the middle range and some in the top range.

The category 'reflection' with the indicator question 'The sentence is descriptive ... reflective' can be detected with substantial reliability. This category is the common denominator of the level models that distinguish between descriptive and reflective texts. The estimate of the reliability of the simple majority vote rater was substantial (see Table 8). All common categories of reflective writing correlated positively with 'reflection' (see Table 10).

Most reliably, the category 'experience' can be detected. The indicator question was 'The writer describes an experience he or she had in the past'. This category was frequently mentioned in the investigated models of reflection (see Table 2). It was also the indicator rated most reliably by the human coders with the simple majority vote approach (see Table 8). It had the highest reliability of the dataset, and a moderately positive correlation with reflection (see Table 10).

The indicator question 'The writer describes his or her feelings' from the category 'feelings' was detected with substantial reliability. It was part of more than half of the investigated models. The reliability of the dataset was almost perfect. The reliability estimate of the simple majority vote raters was substantial (see Table 8), and had a strong positive correlation with reflection.

The category 'personal' was captured with the indicator question 'The writer describes his or her beliefs'. This indicator was detected with substantial reliability, and it was mentioned in most models of reflective writing. The dataset reliability was almost perfect; it had substantial rater reliability and was moderately correlated with reflection.

The indicator question 'The writer recognises difficulties/problems' from the category 'critical stance' was detected with moderate reliability. The reliability of the dataset was again almost perfect. All models of reflective writing contained categories

related to 'critical stance'. The indicator question was rated with substantial reliability by the raters, and had a moderate rank correlation with reflection.

'The writer takes into account another perspective' was the indicator question for the category 'perspective'. The reliability of the detection model was moderate, and its dataset reliability was substantial. This was part of many of the models of written reflection. The indicator question had the lowest rater and dataset reliability, and its correlation with reflection was weak.

The 'outcome' category was part of most models of reflection. Two indicators were developed: the first was 'The writer intends to do something', and its detectability was substantial. The dataset had an almost perfect reliability; it had substantial rater reliability and a weak correlation with reflection. The second indicator was 'The writer has learned something', which was predicted with substantial reliability. The dataset had almost perfect reliability, whereas the rater reliability and correlation with reflection was moderate.

The top performing algorithms varied from dataset to dataset, which indicates that there is not a single best algorithm to detect reflection. Both the SVM and Random Forest showed a good overall performance over many datasets. The tree-based and rule-based algorithms did not reach the reliability of the top-performing machine learning algorithms, but they still had substantial reliability.

7

CONCLUSION AND FUTURE RESEARCH

This chapter summarises the intention of this research and discusses the research questions in the context of this investigation in Section 7.1 'Research questions'. This body of work made several contributions to research, which are summarised in Section 7.2 'Contributions'. Section 7.3 'Limitations' discusses the limitations of the conducted research, and Section 7.5 'Concluding remarks' shows several areas of future investigation that became possible with the findings of this thesis.

7.1 RESEARCH QUESTIONS

Chapter 1 'INTRODUCTION' outlined the importance of reflection for education. The significance of reflection is well established and recognised at the highest level of educational policy, for example by the QAA, the OECD, the European commission, and the U.S. Department of Education.

An important educational practice is reflective writing. Despite its importance, the automated detection of reflection in texts has not been widely researched. Research has explored dictionary-based approaches that map key words to categories associated with reflective thinking. It has employed systems that allow to manually construct rules that detect text segments related to discourse, and machine learning algorithms have been explored to classify text according to several thinking skills, for example, the cognitive presence model of Garrison et al. (2001) or the framework of argumentative

knowledge construction of Weinberger and Fischer (2006) (see Section 3.2 'Related automated methods').

The main aim of this thesis has been to investigate the potential of automating the detection of reflection in text. It specifically investigated whether machine learning can be used to automatically detect reflection in text segments.

Two research questions have been investigated. The first research question was **Q1: Can machine learning algorithms be used to distinguish between descriptive and reflective text segments?** The second research question was **Q2: Can machine learning algorithms be used to detect common categories of reflective writing?**

Both research questions stem directly from the research of the manual content analysis of reflective writing. They investigate key features of the analysis of reflective writing. Section 2.2 'Models to analyse written reflection' brought together an extensive list of models used to analyse reflective writing. The point was made that these models exhibit two qualities, which were described as the quality of depth and the quality of breadth. This research addressed both qualities of reflective writing models.

The first research question investigated the depth or level dimension of reflection models. As described there, models analysed reflective writing regarding several levels of reflection. Their common denominator was that a reflective writing can be characterised as reflective or descriptive/non reflective. This distinction was captured in the first research question.

The second research question investigated the breadth quality of reflective writing. These are the categories that are associated with reflection and are frequently analysed in reflective writings. Section 2.3.1 'Evidencing common categories of reflection' showed that there are six categories that are frequently part of the analysis of reflective writing. These were the test candidate categories for the evaluation of the second research question. Furthermore, Section 5.7.4 'Reliability' showed that these

common categories of reflection can be reliably annotated. Section 5.7.5 'Validity' showed that these common categories not only have face validity, but also correlated positively with reflection. This corroborates their validity with empirical evidence.

The datasets that were used to evaluate both research questions were vetted with a strict quality standard. This additional step was taken in order to strengthen the validity of the datasets. They are only formed by sentences that received substantial support to belong to one of the classes of the categories.

Based on the guiding evaluation criteria of reliability and validity, this research took several measures to ensure the objectivity of the research design (see Section 4.2 'Evaluation criteria and metrics'). Section 4.6.2 'Research design' outlined in detail the steps that were taken to evaluate the machine learning algorithms on the problem of reflection detection. This was continued in the description of the concrete implementation of the data generation process (see Chapter 5 'DATASET GENERATION'). The researcher did not take part in the data generation process and the raters of the sentences did not know the writers of the texts. The raters coded the data independently from each other based on explicit coding instructions. The raters received as training the same test questions from a pool of test questions. Although this seems a matter of course, Poldner et al. (2012, p. 32) remarked that only one of the reviewed 18 papers about the content analysis of reflective writing provided detailed information about the coding instrument.

In addition to these measures, the evaluation was conducted under standardised conditions. This concerned the dataset generation, data preprocessing, with feature selection and construction steps, division of datasets into training and test data, training dataset preparation with the original class distribution and random oversampled training dataset, model selection, based on the same resampling strategy and model tuning, and the model assessment step. The machine learning algorithms

were not individually optimised. This ensured that their differences can be attributed to the algorithms and not to individual optimisation strategies.

Compared to the research about automated methods on related constructs of reflection (see Section 3.2.3 'Machine learning approaches'), this research extended the amount of investigated machine learning algorithms. Compared to the research of the manual content analysis of reflective writing (see Section 3.1 Manual methods to detect reflection) and the automated methods, this research reported the reliability of the automated methods with several reliability measures, all of them found in the related research and an additional reliability measurement.

This is the context in which both research questions have been investigated.

The first research question was **Q1: Can machine learning algorithms be used to distinguish between descriptive and reflective text segments?**

The findings of this thesis strongly suggest that machine learning algorithms **can** be used to distinguish between descriptive and reflective sentences.

The top performing machine learning algorithms had a Cohen's κ of 0.7 and a per cent agreement of 0.89. They came close to the top bracket of the per cent agreement reported by the research of manual content analysis of reflective writing and were reliable according to several benchmarks. In addition, these reliability values are comparable to the Cohen's κ values reported in the research on automated methods of related concepts of reflection (see Section 3.2.4 Automated methods performance).

Furthermore, this result was founded on the investigation of three types of machine learning algorithms. They were: 'I1: Can tree-based machine learning algorithms detect the difference between descriptive and reflective texts segments?', 'I2: Can rule-based machine learning algorithms detect the difference between descriptive and reflective text segments?', and 'I3: Can high performance machine learning algorithms detect the difference between descriptive and reflective text segments?'

This approach provided a comparative overview of different types of machine learning algorithms. It was shown that the third line of investigation achieved the best reliability, followed by the rule-based and the tree-based machine learning algorithms. While the latter two had about the same performance, the high performing algorithms showed better overall performance regarding the reliability measures, and the AUC.

The second research question was **Q2: Can machine learning algorithms be used to detect common categories of reflective writing?**

The findings of this thesis strongly suggest that machine learning algorithms **can** be used to detect common categories of reflection.

Compared with the previous research question that evaluated three types of machine learning algorithms on the same dataset, this research question was evaluated with the same set of high performing machine learning algorithms on different datasets.

All categories have a per cent agreement that is in the top or middle bracket of the agreement values reported in research about the manual content analysis of reflective writing. None of the categories is lower than the middle bracket.

The most reliable category is 'description of an experience' with a Cohen's κ of 0.83. Its reliability is in the top bracket of the per cent agreement of the reported agreement values of the content analysis of reflection. It is also in the top of all reported benchmarks. Compared to the reliability values reported by the machine learning algorithms of related research, this category outperforms all of them.

The category 'feelings' has a Cohen's κ of 0.73. Its per cent agreement is in the middle bracket of the content analysis benchmark. It is reliable according to all reliability benchmarks. Its κ is close to the top value reported in the research of automated methods.

The indicator 'intention' and 'learning' of the category 'outcome' has a Cohen's κ of 0.71 and 0.63 respectively. 'Intention' is in the top bracket of the human benchmark and also reliable according to all other reported benchmarks. 'Learning' has a per cent

agreement of 0.83, which is in the middle bracket of the reported agreement values of the manual analysis of reflective writing. It is reliable according to most benchmarks and the κ values of 'intention' are in the top area of κ values reported by the automated methods or related research.

The category 'beliefs' shows a Cohen's κ of 0.66 and a per cent agreement of 0.83. It is reliable according to most benchmarks and in the middle bracket of the human coders analysing reflective writings. It is in the area of the reported machine learning algorithms that detect related constructs.

The category 'critical stance' with the indicator 'difficulties' achieved a Cohen's κ of 0.6 and a per cent agreement of 0.80. It has a moderate reliability, but is close to substantial reliability.

The category 'perspective' has a Cohen's κ of 0.55, but a per cent agreement of 0.88, which brings it close to the top bracket of the human benchmark.

Overall, these findings strongly suggest that machine learning can be used to detect both, reflection and its common categories.

Substantial effort went into the generation of the annotations in order to produce datasets of reasonable size necessary for machine learning (see Chapter 5 'DATASET GENERATION'). As it is often with supervised machine learning for text classification, most of the time is spent designing and conducting the annotation task. This thesis provided a full account of both the data generation process and the evaluation of the machine learning algorithms with regard to their reliability in detecting reflection.

7.2 CONTRIBUTIONS

This thesis presented work that investigated the potential of automating the detection of reflection in text using machine learning. The main contribution of this thesis was the standardised evaluation of several machine learning algorithms on the problem of

the automated assignment of reflection-relevant labels to text segments. In detail, the thesis contributed with a standardised evaluation of tree-based, rule-based, and high performing machine learning algorithms to the problem of reflection detection. Furthermore, it contributed with an evaluation of machine learning algorithms in order to detect common categories of reflective writing. Several state-of-the-art machine learning models were trained and assessed. The evaluation provided a comparative overview of the potential of machine learning algorithms to automatically detect reflection.

Section 2.2 'Models to analyse written reflection' provided an extensive overview of the models used to assess reflective writings. There, we showed that research on reflective writing proposed many different models to assess such writing (see Table 1). The models were analysed with regard to their commonalities. The results of the synthesis of the models of reflective writing were a set of common categories of reflective writing, where each common category represents frequently found constituents of the models of reflective writing. The categories are summarised in Section 2.3.2 'Common reflection categories'. Evaluation of the categories showed that they can be used to reliably annotate datasets (see Section 5.7.4 'Reliability'), and furthermore, all indicators of the common categories of reflection are positively correlated with the indicator reflection (see Section 5.7.5 'Validity').

Furthermore, the literature on the analysis of reflective writing was prepared with regard to reliability measures. Section 3.1.4 'Manual reflection detection performance' provided an extensive overview on the reliability of raters to analyse written reflection. This overview provided detailed information on the inter-rater reliability of analysing reflective writing.

This thesis also contributed to the research on crowdsourced text annotation. Supervised machine learning requires large annotated datasets. Crowdsourcing the annotation task allowed to divide the coding work over many participants, and thus

scaled well for this large dataset. The contribution made here was to show that crowdsourcing can be used to annotate sentences with regard to reflection indicators. Furthermore, we showed that by aggregating the ratings, the coding process can reach suitable inter-rater reliability (see Section 5.7.4 'Reliability'). Machine learning is widely used in areas of education where sufficient annotated datasets are available. The research on e-assessment, for example, makes use of large text collections of graded essays. In many areas of research, these large annotated text collections are not readily available. The crowdsourced annotation approach shown here can help lower this barrier of using machine learning in order to analyse learning expressed in text.

This research also contributed to practice. Automated methods for analysing reflection and learning in general are important for providers of large online learning environments. The description of the requirements for sample size and the annotation process allows assessing the effort required to generate one's datasets for the development of machine learning models to detect reflection. A detailed description of the performance of the machine learning models allows evaluating their suitability for specific applications.

7.3 LIMITATIONS

The main focus of this thesis was to investigate whether *text segments* can be analysed automatically using machine learning algorithms to detect the presence (or absence) of key reflection elements (see Section 1.1 'Research questions'). Research on the manual content analysis of reflection indicated that shorter text segments are used frequently as analysis units (see Section 3.1.4 'Manual reflection detection performance'). This was one of the reasons for choosing sentences as analysis units Section 4.6.1 'Dataset generation process'. The entire text as an analysis unit was also mentioned frequently

in the literature. The applicability of machine learning algorithms to detect reflection at the text level was not investigated, and this was expressed in the research question.

The study investigated only text written in English; other languages were not explored. The hope is that this research can inspire further research to evaluate the potential of automated detection of reflection across languages.

7.4 FUTURE RESEARCH

The automated detection of reflection opens several lines of investigation. Four of them are outlined next. They are the extension of this research to the reflection detection of other languages, the extension to spoken language, the application of reflection detection in research, and the long term ambition – the automated assessment of written reflection.

One topic was already mentioned in the limitations of this chapter. It is the question: to which degree can reflection be captured in other languages with machine learning? This research showed that the singular first person pronoun was an important feature to distinguish reflective from descriptive sentences. In written German the use of the singular first person pronoun is relatively rare as school teaches an objective, not subjective writing style. Other language specific features would open interesting perspectives for the automated detection of reflection. Another research challenge are languages that do not have word boundaries. For example, written Chinese is a continuous stream of characters. But also the scriptio continua writing style of old Greek or Latin does not use word boundaries. Word boundaries were important for this research as they were used to convert sentences into features. These features were necessary to train the machine learning models.

Detection of reflection in spoken language: For this thesis, the main text source were writings. Written language is one form of expressing reflection. Another form is

the spoken language. The research design developed in this thesis is seen as applicable not only to the written language, but also to spoken language that has been transcribed. The reflection detection model synthesised from the research of written reflection is sufficiently general to be adapted to the context of the spoken language. The main reason for this is that most models of reflective writing, from which the model for this thesis was synthesised, are based on the reflective literature of authors of great acceptance (see Section 2.1 Definitions of reflection and Section 2.2 'Models to analyse written reflection'). They are bound to reflective thinking, but not the medium. The investigation of spoken language with automated methods comes a step closer to researching reflection-in-action (Schön, 1983, 1987). Reflection-in-action is mostly an inner thinking process. The person is stepping back from a problem and reflects on the situation in situ. These reflections can be captured in writings, but this happens after the event and thus it is part of the reflection-on-action dimension. Researching spoken language opens the potential to capture the utterances of a person that reflects in action. Considering the amount of utterances made during a day, the automated detection of reflection can help to identify moments of reflection-in-action. Further, reflective writing is mostly an activity of an individual. The research on spoken language allows to study the utterances of groups and thus allows to determine reflective utterances in the context of the group interactions. The automated detection of reflection can help to detect the moment when a reflection is triggered.

Application of reflection detection in research: The immediate application of the detection of reflection is seen as a research tool to analyse text for reflection on a large scale. The effort that has to be put into the manual analysis of reflective writing limits its scope to relatively small sample sizes (see the column 'number of texts' in Table 3 of Section 3.1.4 'Manual reflection detection performance'). The effort to assess reflective writing often restricted research to single group investigations measured at a

single point in time. Automated detection allows repeatedly analysing the writings of massive text collections. In addition to analysing the frequencies of reflection categories for each text, the strategies used by the researchers to analyse written reflection can be used to assign categories to entire texts. (see Section 3.1.2 'Relationship between analysis units and reflection categories'). This opens the potential to research many of the questions investigated by the researchers of the manual analysis of reflective writing on large scale. In addition, it allows to investigate in online learning systems which of the learning activities lead to reflective writing. Reflection detection can then be used to inform learning design.

Automated assessment of reflection: One long-term ambition leading from this research is the automated assessment of reflection. Although the automated detection of categories of reflection is a big part of the assessment of reflective writings, assessment, however, is more than the detection of reflection: it includes other dimensions, for example, feedback mechanisms and educational assessment quality standards. This is a wider theme that needs to be researched.

Automated methods for detecting reflection can have a unique benefit over manual methods. Reflection can be an extremely personal matter, and therefore, people might prefer not to share their reflection with others. An automated system could provide meaningful feedback while removing the fear of being personally judged.

The manual assessment of writings is a time-consuming task. In courses with many students, it might not be possible to provide feedback on all the writings precisely because, assessment is time-consuming and expensive. One way of overcoming this workload is to assess an aggregated report based on several reflective writings (called a reflective account). However, by emphasising that frequent writing helps learning, automated systems can deliver feedback instantly on each reflective writing. The automated assessment can provide additional and on-time feedback along with the feedback provided later by an expert.

Furthermore, the information gathered by an automated system can serve teachers as a second opinion on their assessment of reflection. The pre-annotated text of the reflection detector can be used as the first iteration that can then be refined through teacher comments. The automated system can serve as another perspective on the matter, which may improve assessment accuracy (Winkler and Clemen, 2004).

A concrete application scenario of this technology for education is, for example, to support learners with automated formative assessment about their reflective writing. A teacher introduces the learners to reflective writing outlining the value proposition of reflective thinking and writing (see Chapter 1 'INTRODUCTION') as well as providing information about common constituents of reflective writing (see Section 2.3.2 'Common reflection categories'). These categories represent concrete learning goals. The teacher outlines that a good reflective writing should contain evidence of these constituents. Then, the learner engages in the reflective writing activity using a system that automatically detects text segments that belong to the common categories of reflection. The learner can check the current draft of the reflective writing with the automated reflective writing assessment program, for example, whether or not all categories of a reflective writing have been covered or whether crucial parts of a reflective writing, for example, the description of the problem was missing. Based on the assistance of the automated system, the learner can modify the current draft and request feedback again at a later stage. The benefit of using such a system is that the learner receives this feedback instantly and continuously and not delayed and less frequent from the teacher as it was the case without this technology.

7.5 CONCLUDING REMARKS

Empirical research at the convergence of educational science and computer science made it possible to investigate this research question. The guiding discipline was educational science, which provided the model for reflection detection, and computer science provided the automated method to detect reflection. Only the synergy between both allowed to study the automated detection of reflection.

A

CO-CITATION ANALYSIS OF RESEARCH ON REFLECTIVE WRITING

The literature on reflection is constantly growing. A co-citation analysis (White, 2011, p. 277) was conducted to identify high impact research papers in the area of reflective writing.

A search in the citation database of Thomson Reuter's 'Web of Knowledge'[1] for the words reflection and writing in the title of papers returned 358 results (February 2012). Research areas like optics, geography, surgery, chemistry, religion, or literature were excluded from the search result. From these articles the references were extracted. The following figure (Figure 12) shows a condensed view of the analysis showing only papers which got cited more than 3 times within the used data set.

The most cited research articles according to this co-citation analysis are listed below. It includes the seminal work of Dewey (1933), as well as the influential work of Schön (1983, 1987); Mezirow (1981, 1990b, 1991); Boyd and Fales (1983); Boud et al. (1985). An early review on the topic of reflection was written by Atkins and Murphy (1993).

Two papers focus on reflective writings using journals (Hahnemann, 1986; Paterson, 1995). Both describe reflective practice in the nursing area, which has a long tradition in reflective practice. The other well researched area is teacher education, which is also reflected in the co-citation analysis with the papers of Zeichner and Liston (1987) and Hatton and Smith (1995).

1 http://apps.webofknowledge.com

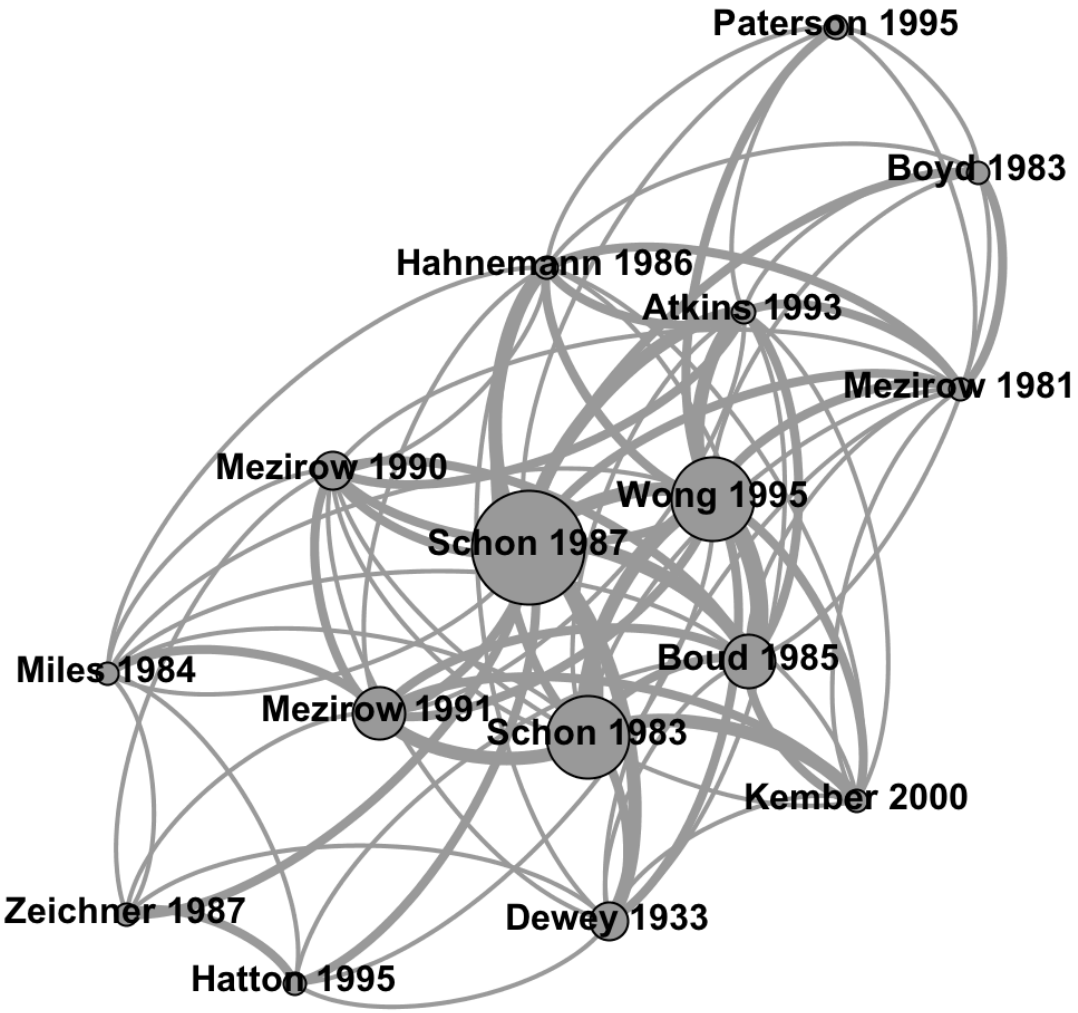

Figure 12: Co-citation analysis of the topic reflective writings

For this thesis the area of assessment of written reflection is especially important, as it researches the method of identifying reflection in texts. The two papers of Wong et al. (1995) and Kember et al. (2000) deal especially with the topic of the assessment of written accounts. The book of Miles and Huberman (1984) gives a general introduction to qualitative methods.

B

MAPPING OF MODELS OF REFLECTION TO COMMON CATEGORIES OF REFLECTION

The following Table 37 shows the mapping of the models of reflection (Table 1) to the common categories of reflection. This is the long version of Table 2. Each cell contains the evidence found in each model that supports its classification into one of the common categories of reflection. See Section 2.3 'Model for reflection detection' for the detailed description of the common categories of reflection, the models, and the mapping process. See Section 2.2 'Models to analyse written reflection' for the description of the models (especially Table 1).

Author(s)	Description of an experience	Feelings	Personal	Critical stance	Perspective	Outcome
Sparks-Langer et al. (1990)	Descriptive language, events	✗	Explanation with personal preference as justification	Explanation, rational	Taking into account the theoretical, and/or contextual perspective, and/or considering principles ethical, moral, or political issues	✗
Wong et al. (1995)	Prior knowledge, feelings, and attitudes	Attending to feelings	Making something one's own, building identity	Re-assessment, finding relationships, testing consistency, and validation	✗	Outcome of reflection (transformation of perspectives, change in behaviour, commitment to action), and appropriation
McCollum (1997)	Description of the event and situation	Implicit	Implicit	Justification and critique	Focus on situation and sensitizing	✗
Kember et al. (1999)	Description of the subject matter	Feelings part of the reflection	Reflection about own perceptions, feelings, and actions, premise reflection (becoming aware of own unconsciously held beliefs)	Assessment of efficacy and critical review of assumptions	✗	Change of perspective, better awareness
Fund et al. (2002)	Description	Concerns coming from feelings	Personal opinion	Critical bridge	Discussion of alternative opinions, analysis of context factors	Insights, better awareness, reaching conclusions
Hamann (2002)	✗	Attention to emotional needs	Grounding reasoning in the knowledge of self	Tentative conclusions, means-end and strategic thinking, reasoning	Taking into account the emotional, cognitive, social, ethical or political context	✗
Pee et al. (2002)	Descriptive type and descriptive reflective type	✗	Discourse with one's self	Reasoning	Exploration of alternatives, socio-political context	✗

Continued on next page

Table 37 Continued from previous page

Author(s)	Description of an experience	Feelings	Personal	Critical stance	Perspective	Outcome
Williams et al. (2002)	Description of the learning event, issue, or situation	Feelings mentioned in three of the five criteria	Implicitly given, as the model focusses on the individual experiences of the participants (their prior knowledge, feelings, and attitudes)	Reasoning, analysis, and verification	✗	Gains new understanding, indicates future behaviour
Boenink et al. (2004)	✗	Doubt, uncertainty	Personal perspective	Balanced approach, evaluation	Taking several perspectives into account	✗
O'Connell and Dyment (2004)	✗	✗	Personal reflection	Analysis, synthesis, and evaluation	✗	Learning outcome in the dimensions of knowledge, comprehension, application, analysis, synthesis, evaluation, transfer, and personal development
Plack et al. (2005)	Return to experience (significant experience)	Attending to feelings	Model centres on the view of the student. See, for example, the category 'return to experience'	Process (describing strategies), premise (exploring and critique), re-evaluation (validating insights)	Premise (seeking multiple perspectives), content (exploring experience from several perspectives)	Reflection-on-action and content (better understanding), Reflection-for-action (planning for the future), re-evaluation (appropriation of new meaning into own way of being)
Ballard (2006)	Technical rationality	Being concerned	Implicit	Clarifying assumptions, addressing consequences and personal biases	✗	✗
Mansvelder-Longayroux (2006); Mansvelder-Longayroux et al. (2007)	Recollection of past experiences	✗	Implicit, as focus is on personal experiences	Evaluation, analysis, critical processing, diagnosis	Comparison of own opinions with other opinions	Examination of progress made so far, lessons-learned, capturing consequences of action

Continued on next page

Table 37 Continued from previous page

Author(s)	Description of an experience	Feelings	Personal	Critical stance	Perspective	Outcome
Plack et al. (2007)	Description of experience	Feelings	Perspective of the writer	Awareness of gaps, analysis, evaluation	Multiple perspectives	Articulation of learning, conclusions, and future actions
Kember et al. (2008)	Personal experiences	✗	Personal experiences	Critical reflection	Taking perspectives into account	Change of perspective, awareness of own beliefs
Wallman et al. (2008)	Description of the course of events	Personal feelings of what was felt and awareness of the function of feeling on own behaviour	Thoughts and feelings about oneself	Analysis of problems, questioning behaviour, analysis of the whole situation	Alternative methods, reinterpretation	Awareness of prejudices, new understanding of the situation for future action
Chamoso and Cáceres (2009)	Description of learning process	✗	Implicitly indicated in the examples supporting the model	Argumentation, justification, conclusion	✗	Involvement, contributing improvements
Lai and Calandra (2010)	Description of concerns of practice	Concern, frustration, excitement	Personal involvement	Inquiry, questioning	Openness to conisderations of others, active search of other perspectives	New insights, Re-framing of perspective, change of practice
Fischer et al. (2011)	Reporting, description of the personal interpretation of experience	Feelings	Awareness of own feelings, judgement process, and beliefs)	Assessment of judgement process, questioning, assumptions	✗	Awareness about the specific personal judgement process
Birney (2012)	Description of context	Emotions	Self-awareness	Analysis, synthesis, issue identification, implications of actions	Multiple perspectives, context	Learning, understanding, change of beliefs, intentions
Wald et al. (2012)	Writing spectrum	Attending to emotions	Presence of author	Awareness of conflict or dilemma, analysis, meaning making	Multiple perspectives	Transformative or confirmatory learning
Mena-Marcos et al. (2013)	✗	✗	Implicit	Analysis, evaluation	✗	Artefact to solve a problem
Poldner et al. (2014)	Description of context	✗	Implicit	Evaluation and justification	✗	Transfer

Continued on next page

Table 37 Continued from previous page

Author(s)	Description of an experience	Feelings	Personal	Critical stance	Perspective	Outcome
Prilla and Renner (2014)	Description of experience	Emotions	Implicit	Awarness of issues, providing a rational/reasons, challenging or supporting assumptions	Alternative perspectives	Solution proposal, learning, intention to change, change

Table 37: Mapping of models to the common categories of reflection

C

SAMPLED TEXT COLLECTION OF THE BAWE CORPUS

The following Table 38 contains all texts from the BAWE corpus text selection process (see Section 5.2). The table describes for each text the taken decision to keep it, delete it, or to prune it, as well as a comment section describing the reason for these decisions. Some of the texts did not come with a description of the title. These are left empty.

Identifier	Title	Decision	Comment
0003b	An urban ethnography of two bookshops in Leamington Spa, November 2004	delete	descriptive
0050c	CDA assignment 2	delete	descriptive
0055b	What does Descartes mean by a 'real distinction' between mind and body? How does his argument for the real distinction work?	delete	descriptive
0056a	Experiences at King Henry V111 School, Spring 2005	keep	
0057a	Must a subject/object-oriented conception of logic admit to the revisability of logic?	delete	descriptive
0065g	A medical student faced with the 3 suicide bombings of Cairo in April 2005	keep	
0077f	A Cross-sectional Analysis of Economic Growth Theory	delete	descriptive
0114g	A medical elective at McMaster University, March 8th-April 26th 2004	delete	descriptive
0116d	The Impact of Drama on High School Students' English Learning and Self-Confidence	delete	descriptive
0152c	Governance Frameworks and a Review of a Firm's Social Responsibilities	delete	descriptive
0165d	CILM Reflective Piece 1	keep	
0169i	Assignment Term One - Learning Diary	keep	
0172c	ESSAY PROJECT: ACTIVE DESKTOP IMPLIMENTATION AT ASTON MARTIN	keep	

0193d	CFS Insurance - Restructuring Project	pruned	
0202k	1. Why is quality important to EHL? (20% of the marks) 2. What are the underlying causes of the quality problems at EHL? (40% of the marks) 3.What steps would you advise Paul Stone to take to improve quality performance at EHL? (40% of the marks)	keep	
0206e	Strategy Learning Review	keep	
0206l	Question 1	delete	descriptive
0212d	Assignment 1	delete	descriptive
0234a	CFS Insurance	pruned	
0234j	What, if anything, has membership of a political community in common with membership of a family? Can this tell us anything about our obligations?	keep	
0235c	If experience is the source of our grasp of the concept of a physical object, then how do we make intelligible to ourselves the idea that such objects may exist unperceived?	delete	descriptive
0237a	Introduction	delete	descriptive
0253c	Reflection on Hofstede's Cultural Dimension	keep	
0253d	Reflection on the Shell's Stakeholder Approach	keep	
0311k	Is the Simplicity of a Theory any Guide to its Truth?	delete	descriptive
0316c	How Do the Media Contribute to Changes in Time and Space?	keep	
0325b	Countable and Uncountable Sets	delete	descriptive
0342b	Design Project Circuits	delete	descriptive
0342c	Reflective piece	keep	
0347g	Reflective Piece - Business Project	keep	
0348c	Individual Reflective Piece	keep	
0354a	Individual Reflective Piece	keep	
0354f	Writing Exercise: What Makes a Professional Engineer?	keep	
0358h	Reflection: Entrepreneurial Activities are always quite interesting and demanding.	keep	
0362b	First Reflective Piece	keep	
0362c	Second Reflective Piece	keep	
0393a	A critical assessment of Derek Parfit's fission example in his 1971 article "Personal Identity".	delete	descriptive

0399a	A discussion regarding the statement that: "The extent to which a monopoly induces economic efficiency depends, among other things, upon factors such as: (i) Vertical integration in the market, (ii) Horizontal market contestability, (iii) Technology, (iv) The role of advertising and (v) Market demand elasticity."	delete	descriptive
0401b	How far is the normalisation of drugs relevant within youth culture?	pruned	
0405c	Participant Observation	keep	
0411c	The right to silence; myth or reality? Discuss	pruned	
0415g	Searching the Chemical Literature	delete	descriptive
0424c	Reflective Piece for Starting and Running a Business	keep	
3006h		delete	descriptive
3006i	Analysis of the Intended Market and Marketing Campaign	delete	descriptive
3011b	Learning Portfolio	keep	
3012a		pruned	
3012d		pruned	
3013f	Managerial Features and Approaches to Learning	keep	
3013g	Personal Development Portfolio for an Entrepeneurial Career	pruned	
3019a	Is There A Good Alternative To The Tripartite Theory Of Knowledge	keep	
3019b	Why Does Anscombe Think We Ought to Abandon the Concepts of Moral Obligation and Duty?	delete	descriptive
3019g	To What Extent Would You Agree with the Claim that God's Existence can be Proved? Use One of the Arguments for the Existence of God to Illustrate your Answer.	delete	descriptive
3026c	Assignment 1 - Cover Letter + CV	delete	descriptive
3027a	Anthropology- Library exercise	delete	descriptive
3032g	Mental Health nursing developments in care	keep	
3034e	Individual reflective report on the process of teams work and team work skills	keep	
3034f	Health promotion in midwifery	keep	
3045a	Marketing Comment	delete	descriptive
3047c	Ethnographic journal entries - A cross culture experience	keep	

3052c	Coursework 4 -Project Title: DirSync	delete	descriptive
3052d	Coursework 1	delete	descriptive
3057b	ASSIGNMENT 1 : Analysis of single text, Construction of another in response, Critical comparison	delete	descriptive
3059a	Responding to Others	keep	
3063a	Behavioural observation report : activity of an adolescent female chimpanzee	delete	descriptive
3063c	Research proposal	delete	descriptive
3064a	2.1: Reflective Statement 2	keep	
3064c	2.1: Reflective Statement 1	keep	
3064d	Professional skills: A multi-professional approach	pruned	
3064e	Health People and Society - Essay Assignment	delete	descriptive
3064g	Individual Reflection on Team Work and Team Working skills	keep	
3069a	Viva Case Study - Deirdre	delete	descriptive
3069b	The responding to others interview write-up	keep	
3076b	Stage 2: Critical analysis of my key personal characteristic - Anxiety	keep	
3076c	A reflective account of a decision made in practice incorporating critical analysis of the evidence that was/could have been used to inform the decision.	keep	
3082b	Reservations Manager of Ed	delete	descriptive
3089d	A book proposal for the crime fiction market	delete	descriptive
3092d	A reflective account on the process of teamwork	keep	
3092e		pruned	
3092f	Reflect on a challenging experience in relation to the practice....	keep	
3092g	Stage 1: reflective journal entry and critical analysis of a personal characteristic/behaviour identified from the reflective journal entry.	pruned	pruned literature
3092h	A reflective account of a decision made in practice incorporating critical analysis of the evidence that was/could have been used to inform the decision.	pruned	took out literature review
3094f		pruned	

3099d	Essay Question 2	keep	
3101b	Self-evaluation report	keep	
3101d	Self-development essay	keep	
3110c	The First Men on Mercury	delete	descriptive
3110d	Post-modern American Poetry: the New York Poets	delete	descriptive
3113a	Professional management experience - Year-out work experience	delete	descriptive
3114a		keep	
3114b	Reflective case study of a family under stress	keep	
3118b	Bridging the gap: A case study describing the experiences of a Chinese student studying on an EAP pre-Masters course at Brookes University	delete	descriptive
3119b	How are social inequalities reflected in what people eat? What positive measures can be taken by health and social care professionals to reduce inequalities in the diets of patients and service users?	pruned	
3125a	Assignment 2 Critical review task	delete	descriptive
3125f	Psycholinguistic reflective review	keep	
3126d	Gender identities assessment	delete	descriptive
3127a	Should ESOL teachers be trusted to create their own approaches to learning?	pruned	
3141a	Reflective paper focusing on the first session interview	keep	
3145b	The realisty of evil makes it impossible for the God of classical theism to exist. Discuss.	delete	descriptive
3147d	To what extent would you agree with the claim that God's existence can be proved?	delete	descriptive
3147g	Analysis of the intended market and marketing campaign	delete	descriptive
3150a	Portfolio 2 : Grammar teaching and learning	keep	
3150b	Portfolio 4 : Skills teaching	delete	descriptive
3150c	Portfolio 5 : Language games and activities	delete	descriptive
3150d	To what extent are students interested in history?	pruned	
3157b	Digitising using Geographical Information Systems	delete	descriptive
6002c	Experiment E: Viscosity Measurement	delete	descriptive
6002d	Experiment F: Surface Tension	delete	descriptive

6005g	STAR REPORT	delete	descriptive
6009a	Discuss how theories of second language learning and teaching can explain certain features of your learning Persian.	delete	descriptive
6018a	Assignment 2 [maels1ia]	delete	descriptive
6020c	LS2ASD Assignment	delete	descriptive
6024a	Motivation, Imput and output in my Persian Learning Experience	pruned	deleted the mostly descriptive input hypothesis section
6028a	Towards an understanding of Second Language Acquisition	pruned	deleted introduction, section 3-3·3, 4·4·3 as they are mostly descriptive
6028c	Towards an Understanding of Syllable and Syllable Structure	pruned	
6048b	Do the frequencies of singular and plural first person pronouns within politicians' speeches reflect the frequencies found in the BNC spoken corpus?	delete	descriptive
6062a	How does Tony Blair's use of the 'I' and 'we' pronouns in his political speeches demonstrate his own and his audiences membership to a particular social group or groups.	delete	descriptive
6062c	How do broadsheet newspapers and online based news sites differ in their use of adjectives in reporting the English nation football team's victories and defeats?	delete	descriptive
6100c	Individual Write-up	keep	
6101a	Computer Vision Assignment: Image Compression	delete	descriptive
6101b	[Linear Algebra for Computer Vision and Robotics I, Robot 2]	delete	descriptive
6101c	The Social, Legal and Ethical Aspects of Computer Science	pruned	
6101f	Compiler	delete	descriptive
6101j	CS3Q2 - Computer Science Interim Report. Package for Electronic Mapping - PDA Version	delete	descriptive
6101k	Practical 1	delete	descriptive
6101l	Assessed Assignment 2 - Lists	pruned	
6102d	[Report for assignment of Virtual reality]	delete	descriptive

6108a	Introductory Programming 1 - Practical 8 Writeup	delete	descriptive
6108b	Internet and Web Technology Integration Report for fictional retailer FilmReel UK Ltd. - Part II: The Systems Response to the Business Requirement	delete	descriptive
6114e	Stokes Drift: the strange effect of waves on water transport	delete	descriptive
6145c	Using your hypothetical case study EITHER explain the relevance of different motivational theories to OR explore the role of communication in the situation described.	delete	descriptive
6153b	Management Case Study: Farm Manager	delete	descriptive
6167e	Valerie Bash is 30 years old and pregnant with her first child. ...	delete	descriptive
6168b	Valerie Bash is 30 years old and pregnant with her first child. ...	delete	descriptive
6169f	Robot football	delete	descriptive
6173c	Second Language Teaching and Learning	delete	descriptive
6174a	'What is the effect of the social variables of class and gender on the way we speak? Is there a link between them?'	delete	descriptive
6174b	Word Association Experiment	delete	descriptive
6189a	Report - multimodality of advert and interview analysis/techniques	delete	descriptive
6203a	REPORT ON THE HUMAN SKELETAL REMAINS	pruned	
6203h	Powder flask	delete	descriptive

Table 38: Relevance sampling of texts

D

RELIABILITY ON INDIVIDUAL LEVEL

This appendix reports the inter-rater reliability and agreement values for a subset of 1000 sentences of the annotation task. They were calculated on individual coder level and not on the aggregated ratings of several coders (see Section 5.7.4 'Reliability').

The inter-rater reliability between the annotators was measured using Krippendorff's α for nominal data, as well as Gwet's AC_1. Additionally, the per cent agreement is reported. These are the values for the whole data set (without any filtering or removal of gamers). The six-point Likert scale was dichotomised.

Indicator	Kripp. α	AC_1	%-agree
The writer describes an experience he or she had in the past	0.45	0.45	0.62
The writer describes his or her feelings	0.33	0.4	0.68
The writer describes his or her beliefs	0.29	0.29	0.64
The writer recognises difficulties/problems	0.38	0.38	0.69
The writer takes into account another perspective	0.21	0.34	0.64
The writer has learned something	0.23	0.27	0.62
The writer intends to do something	0.32	0.63	0.76
The sentence is descriptive ... reflective	0.23	0.28	0.63

Table 39: Reliability and agreement values on individual level

TASK DESIGN

Title of the task: Categorise sentences into 8 categories and write an explanation

Instructions: The task is about categorising sentences. You will have to rate each sentence according to eight questions. Afterwards, write a short explanation to justify your choice of the last question (min. 20 characters).

One example should guide you through the task:

Text: "I need to do some thinking about how to act next time to prevent this interruption from happening or to deal with the situation when she starts".

First you will rate the sentence according to eight criteria (six-point scale). The first six questions ask how much you agree or disagree with the question. The last question asks, if the sentence is descriptive or reflective.

It is important to keep in mind that the sentence has to speak for itself. Do not interpret too much into the sentence.

Afterwards write a short text to explain your choice for the last question. For example: #The interruption should not have happened. He intends to put more thought into dealing with such situations. It is a personal sentence referring to an important experience of the writer. The writer intends to find a solution.

Be aware that your explanation has to start with "#" (see example). If your explanation does not start with a #, it will not be accepted. This is to ensure, that you have read the instructions.

Make sure that your explanation is unique, as every text is unique. I cannot accept

copy-and-paste explanations. Your explanation has to be in English. Use full sentences. In case your explanations are too similar or too dissimilar to the content of the page, you will receive a message. If this happens, please change your answer. Do the same if you receive the feedback that your answer does not look right.

Several test cases are embedded in the task.

Categorise the following text:

<Each individual sentence was inserted here>

The writer describes an experience he or she had in the past *(required)*

disagree □□□□□□ agree

<Each individual sentence was inserted here>

The writer describes his or her feelings *(required)*

disagree □□□□□□ agree

<Each individual sentence was inserted here>

The writer recognises difficulties/problems *(required)*

disagree □□□□□□ agree

<Each individual sentence was inserted here>

The writer describes his or her beliefs *(required)*

disagree □□□□□□ agree

<Each individual sentence was inserted here>

The writer takes into account a different perspective *(required)*

disagree □□□□□□ agree

<Each individual sentence was inserted here>

The writer has learned something *(required)*

disagree □□□□□□ agree

<Each individual sentence was inserted here>

The writer intends to do something *(required)*

disagree □□□□□□ agree

<Each individual sentence was inserted here>

The sentence is *(required)*

descriptive □□□□□□ reflective

<Each individual sentence was inserted here>

Give a short explanation to justify your choice of why you think the sentence is descriptive or reflective:

<text field>

Instruction: Your explanation has to be at least 20 characters long. As every text is unique your answer should be unique as well. Do not provide the same reason for every text.

F

EXAMPLES OF THE DATASETS

The following tables lists sentences that were randomly selected from the datasets of the indicators **Experience**, **Feelings**, **Beliefs**, **Difficulties**, **Perspective**, **Intention**, **Learning**, and **Reflection**. Each table shows for each sentence the number of ratings indicating the presence (Pres.) or absence (Abs.) of the indicator. As an example, the first sentence of the following table was rated seven times that experience is present, and one time that the sentence does not express experience. The last column shows the document information (Doc.). Sentences from the BAWE are referenced by their unique identifier and text that were taken from the exemplary samples of reflective texts are marked with the family name of the author. The first five sentences of each table are instances expressing the indicator and were labelled as the positive class, and the last five sentences are instances that do not express the indicator according to their ratings and were labelled as the negative class. The labels were used to supervise the training of the machine learning algorithms.

More information about the process that generated the datasets from which these sentences are taken can be found in Chapter 5 'DATASET GENERATION'.

Examples of the dataset Experience 'The writer describes an experience he or she had in the past':

Sentence	Pres.	Abs.	Doc.
With hindsight I believe that if we had focussed more on promoting the niche market -especially after identifying a virtually identical competitor!	7	1	0342c

Continued ...

Sentence	Pres.	Abs.	Doc.
Because I felt uncomfortable that other nurses regarded Mary's attendance as inappropriate it made me think about my attitude to those patients I had regarded in this way and it helped me understand the effects that this could have on the care I give.	7	1	3092e
For me this stage was about getting to know what the others expected of me and what I expected of the group.	7	1	3092d
One customer was really nice when this happened because she had seen me being taught what to do.	7	0	Moon
This proved not to be enough and I collected phone numbers to call and see why they did not attend.	8	1	0342c
Simmons argues that what he calls "positional duties" are distinct from "the moral requirement to fulfil positional duties".	1	5	0234j
I felt that could be in the form of a midwife, breastfeeding councilor, or breastfeeding manuals that had been tried and tested by women.	1	6	3034f
In fact, according to McKinsey and Company -1989-, "the feasible improvement in gross margin on sales through improved quality performance was rated at an average of 17%" -Dale 2001:15-.	1	7	0202k
To aid this, I might ask them to confidentially write a list of their preferred partners.	1	8	3127a
Receiving positive feedback should be used as an opportunity to improve self-esteem while negative feedback should not be allowed to reduce esteem as negatives can always be turned into positives -Ward.	1	8	3092g

Table 40: Example sentences of the dataset Experience

Examples of the dataset Feelings 'The writer describes his or her feelings':

Sentence	Pres.	Abs.	Doc.
My choice of research topic is influenced by my values and interests and I chose to study the normalisation of drugs as I find the subject fascinating.	5	1	0401b
Although, the final business plan covered and explained everything in detail I was not satisfied with it as I always found something missing and I tried to discuss with my group but they didn't want to think in so much detail.	8	0	0362c
If I were to take part in another team assignment or have the opportunity in practice I will be more willing to voice my opinions and be more confident about my ability to delegate tasks and take more responsibility.	6	1	3034e
I feel really privileged about that though, because in this small way, I am able to almost have an experience of what real entrepreneurs go through.	6	0	0424c
When I started work at Cowley Manor, I was expecting there to be high expectations of me.	8	0	3101d

Continued ...

Sentence	Pres.	Abs.	Doc.
- Flawless Delivery- Development and deployment by January 5 th- Design and development within budget of 15k.	1	6	0172c
My mentor and I often asked Elle how she was feeling, keeping in mind that she had a tendency to hide her pain from us so that we could keep on top of her pain management.	1	8	3114b
This model explains the process of listening to each other, interpreting what is being said and asked, and then taking action and doing what needs to be done.	0	8	3114b
Becky and I then sat with Joseph and explained calmly that it wouldn't be possible to see a plastic surgeon but that we would do everything we could to insure a minimal scar and that we would be able to treat him shortly.	0	8	3092f
When Susie was in pain and crying, she would get a great deal of attention and close cuddling with her mother.	1	9	3114a

Table 41: Example sentences of the dataset Feelings

Examples of the dataset Beliefs 'The writer describes his or her beliefs':

Sentence	Pres.	Abs.	Doc.
Personally I consider output as a useful tool for retention, motivation and mastery of L2.	6	0	6028a
I believe it would be improved if we can concentrate a question arisen from the topic area and discuss it in the discussion.	7	0	3125f
Because I tried, and failed, to assert myself this placement left me feeling wholly inadequate and unvalued, resulting in a complete lack of confidence in myself and my abilities.	8	0	3092g
Starting the day about 3-4 hours earlier than I was used to, was probably the best plan I had for long time.	6	0	3011b
I think the reality of peer assessments -offering either a harsh rebuttal or encouragement to continue- gave us all incentive to work hard during the break.	10	0	0348c
She responded very positively to all health promotion that was offered as well as researching information by herself, health promotion was easier to carry out on this occasion as it is often dependent on public participation -Ashton and Seymour 1998-.	1	6	3034f
Levin -2005- adds to this say that completer/finishers try to prevent mistakes and omissions, search out aspects of the task that need a lot of attention, maintain urgency and finalise tasks with attention to detail.	1	7	3092d
Group X16 of module ES2A6 comprised three men and two women, all British except for one of the women who was from China.	1	6	0354a
The course is building on the skill set that I have been developing over the past two years; more specifically presentation and teamwork.	1	5	0169i
Next day it was reported in the paper that the child had been taken to hospital seriously ill - very seriously ill.	0	9	Moon

Continued ...

Sentence	Pres.	Abs.	Doc.

Table 42: Example sentences of the dataset Beliefs

Examples of the dataset Difficulties 'The writer recognises difficulties/problems':

Sentence	Pres.	Abs.	Doc.
I assumed that I could just re-engage - but I did not realize that the game had changed.	7	1	Moon
I found the business plan to be very challenging on the first site but I had many learning outcomes and various practical experiences of setting up and running a business.	7	1	0362c
Perhaps I spent a bit too long researching their experience and making sure it was all feasible and factually exacting, but forging backgrounds for these two high flyers made me consider that I need to be doing more in pursuit of personal excellence.	5	0	0348c
Jane was also unaware that she could write her specific wants and needs in a birth plan which should have been formulated with her midwife -DOH, 1993- this would have made her intention to breastfeed clear.	7	1	3034f
I managed to provide some good comments but at the same time realized that I missed out some key points.	9	0	3011b
The motivation was basically to make phone calls absolutely free after one time investment.	1	7	0362b
Further, truly "world class" firms are based on having sufficient concepts, competence and connections.	1	5	0169i
Independence and creativity are both highly emphasized in appendix 8 and also represent entrepreneurial traits.	0	9	3013g
In the modern rewriting, she wears the clothes of a supervisor of a holiday camp with her orange cap, her shirt and her banner bearer.	0	10	3012d
3 In Images of Strategy -pp.356-382- edited by Cummings, S. & Wilson, D. I was also a little surprised to see Smith talk excitedly about EVA/SHV as if they were recently discovered magical formulae for success.	1	9	0206e

Table 43: Example sentences of the dataset Difficulties

Examples of the dataset Perspective 'The writer takes into account another perspective':

Sentence	Pres.	Abs.	Doc.
I perceived Mrs Stacy as appreciating this and found it helpful as she said thank you when I told her the next thing that will be done and at what time, her expression gave me the impression that giving this information over was a positive thing, therefore I continued to do so.	7	1	3114b

Continued ...

Sentence	Pres.	Abs.	Doc.
She goes on to explain that patients labelled in this way often feel they are 'told off' by almost every member of staff and punished by the negative attitudes shown towards them and by being made to wait excessively -Walsh, 2000-.	7	1	3092e
Until that time I probably have ignored or overseen many issues, which might have disappointed some of the students I lived with or made them uncomfortable.	6	1	3011b
For this reason, I will try and be aware of all the issues that might affect an individual and how they behave, for example, whether they are male/ female, old/ young, disabled, elderly, from an ethnic minority or gay.	8	0	3119b
Within our group, we had to be understanding of each others ideas and attitudes, and recognise that we may have different ways of going about things or communicating.	9	0	3064g
After week 4 in the restaurant, we were split into teams to look after tables.	0	8	3101b
She agreed to come back - but came back to see Geoff, the senior partner, still complaining about the shoulder.	1	5	Moon
Luckily we were all dedicated within our group and had similar styles of working which enabled us to consistently be organised throughout.	1	8	3069b
However, one can easily tell the weaknesses of certain types of input such as the one provided in Grammar Translation method.	1	7	6028a
It would not be practical to have a person in the 50's applying for this position, because the potential career is seen as too short and costs for the training too high.	1	5	3011b

Table 44: Example sentences of the dataset Perspective

Examples of the dataset Intention 'The writer intends to do something':

Sentence	Pres.	Abs.	Doc.
I will continue my good teamwork, as I feel I am good at it.	8	0	3101b
For the future, I will try to engage with the cultures I am meeting in group work situations or at work beforehand, in order to be better prepared.	9	1	3011b
I discussed my assignment with Jane and received consent that she was happy to be followed up outside of her next antenatal visit -which would be with her GP at 36 weeks-.	6	1	3034f
Furthermore I have decided to narrow the scope of my activities so I can dedicate more time and commitment to each activity.	10	0	0348c
The topics are very well chosen and the activities and language games will be very useful for my later teacher practice in Germany.	6	1	3150a
The second task was to work on the descriptions of the streets of Saffron Hill.	2	12	3012d

Continued ...

Sentence	Pres.	Abs.	Doc.
The lower individualism in Hong Kong means people in the society are born into extended families and collectives where everyone takes responsibility for fellow members of their group -Hofstede, 2003-.	0	6	3047c
Mentors need to create a relationship that encourages these beliefs for students to feel confident enough to be assertive.	0	7	3092g
Kaizen charges everyone in the organization with improving everything in the company, not just quality.	1	9	0202k
For example, it has been advised that talking behind the computer -or similar- disconnects the speaker with the audience as it creates some what of a barrier between them and therefore dilutes the attention to the speaker.	1	9	0165d

Table 45: Example sentences of the dataset Intention

Examples of the dataset Learning 'The writer has learned something':

Sentence	Pres.	Abs.	Doc.
Thus, through a brief analysis of the learning strategies that I employed in order to approach this daunting task, the errors that I made and the theories behind them I will present how the lack of these three factors affected my progress in the learning process.	8	0	6024a
I can now see how this is far better when you have an inexperienced -and largely unmotivated- team, and will certainly go straight for a structured methodology next time I am in control of a team.	9	0	6100c
Although I felt I could call on and use this training and experience as much of it would apply to face to face situations the prospect of dealing with aggression in person made me nervous as I thought of it as more personally threatening.	6	1	3092f
To conclude, the amount of interaction with native speakers of the language and with other students was too inadequate to practise and check if my performance was correct.	6	1	6024a
Several of my colleagues told me afterwards that Mrs Shaw always steps in to answer questions like that and they commented that I handled her intrusion well.	6	1	Moon
However, it is my opinion that Shell's decision to adopt a stakeholder management strategy is a knee-jerk reaction which arose out of the Brent Spar and Nigeria controversies, respectively.	0	6	0253d
Facts that are used should be correct and avoid making statements that go beyond the facts and might therefore be challenged Empathy - try to be courteous and friendly, however angry we may feel, try to control the emotions and at least remain calm.	1	6	3059a

Continued ...

Sentence	Pres.	Abs.	Doc.
They should also make sure they include all the relevant information and don't try to break the news gently by not making it sound exactly how it is just to ensure their own emotions are satisfied.	1	7	3059a
The media provide an alternative, easier path for accessing a huge amount of information and options that were previously remote, as noted by McLuhan -1967- in his description of the "global village."	1	6	0316c
I started by e-mailing each member of the group to remind them of the meeting.	0	10	0342c

Table 46: Example sentences of the dataset Learning

Examples of the dataset Reflection. The indicator question was: 'The sentence is descriptive ... reflective':

Sentence	Pres.	Abs.	Doc.
On the other hand, I found myself as a stranger in my hometown, this was really true when I have gone to university life while I needed to take the university bus to campus everyday.	10	0	3047c
I need to tell her honestly about the tutorial, the feedback and my disappointment in myself.	6	1	Moon
I was immediately embarrassed by my callous attitude especially when so many people had died and were injured.	6	0	0065g
From a nursing point of view it felt like there was not a great deal that we could do to minimize the stress of having two children in separate places, as explained when describing this possible stressor I attempted to talk with Mrs Stacy about this, however, I feel that a family member or someone who would understand how she felt more would have been a more appropriate person to talk to her about this.	5	1	3114b
Finally I believe that throughout these weeks I have learned some interesting issues about interactive skills and cross-cultural communications.	8	1	3011b
The report goes further by saying that "agitation, delirium, confusion, pain, uncontrolled body movement, hypoxia, faecal impaction and acute urinary retention .	1	6	3076c
This week we are performing some mock appraisal interviews in class, where I will participate as an interviewee and an observer.	0	9	3011b
I will begin by giving some background information on the family, I will then go on to identify the various stressors and explain how the framework can be applied.	1	9	3114b
Hughes states that bed-rails should be avoided due to the risk of injury caused when the patient climbs over them and falls to the floor.	1	5	3076c

Continued ...

Sentence	Pres.	Abs.	Doc.
Dickens mentions it because it was a "local landamark", however, today, it is only the building of a cooperative bank.	1	9	3012d

Table 47: Example sentences of the dataset Reflection

BIBLIOGRAPHY

Abou Baker El-Dib, M. (2007). Levels of reflection in action research. An overview and an assessment tool. *Teaching and Teacher Education*, 23(1):24–35.

Aggarwal, C. C. and Zhai, C. (2012). A Survey of Text Classification Algorithms. In Aggarwal, C. C. and Zhai, C., editors, *Mining Text Data*, pages 163–222. Springer US.

Alden Rivers, B., Whitelock, D., Richardson, J. T. E., Field, D., and Pulman, S. (2014). Functional, Frustrating and Full of Potential: Learners' Experiences of a Prototype for Automated Essay Feedback. In Kalz, M. and Ras, E., editors, *Computer Assisted Assessment. Research into E-Assessment*, number 439 in Communications in Computer and Information Science, pages 40–52. Springer International Publishing.

Alonso, O. and Mizzaro, S. (2012). Using crowdsourcing for TREC relevance assessment. *Information Processing & Management*, 48(6):1053–1066.

Andersen, N. B., O'Neill, L., Gormsen, L. K., Hvidberg, L., and Morcke, A. M. (2014). A validation study of the psychometric properties of the Groningen Reflection Ability Scale. *BMC Medical Education*, 14(1):214.

Anderson, T. and Dron, J. (2010). Three generations of distance education pedagogy. *The International Review of Research in Open and Distance Learning*, 12(3):80–97.

Aphinyanaphongs, Y., Fu, L. D., Li, Z., Peskin, E. R., Efstathiadis, E., Aliferis, C. F., and Statnikov, A. (2014). A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization. *Journal of the Association for Information Science and Technology*.

Association, A. P., Association, A. E. R., and Education, N. C. o. M. i. (1954). *Technical recommendations for psychological tests and diagnostic techniques*, volume 51. American Psychological Association.

Association, A. P., Association, A. E. R., and Education, N. C. o. M. i. (1993). *Standards for Educational and Psychological Testing: Guidelines for One of the Most Important Contributions of Behavioral Science*. American Psychological Association, rev sub edition.

Atkins, S. and Murphy, K. (1993). Reflection: a review of the literature. *Journal of Advanced Nursing*, 18(8):1188–1192.

Attali, Y. and Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning and Assessment*, 4(3).

Axelsson, M. W. (2000). USE-the Uppsala Student English corpus: an instrument for needs analysis. *ICAME journal*, 24:155–157.

Badger, J. (2010). Assessing reflective thinking: pre-service teachers' and professors' perceptions of an oral examination. *Assessment in Education: Principles, Policy & Practice*, 17(1):77–89.

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. P., and Yilmaz, E. (2008). Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 667–674, New York, NY, USA. ACM.

Bain, J., Ballantyne, R., Packer, J., and Mills, C. (1999). Using Journal Writing to Enhance Student Teachers' Reflectivity During Field Experience Placements. *Teachers and Teaching*, 5:51–73.

Bain, J. D., Mills, C., Ballantyne, R., and Packer, J. (2002). Developing Reflection on Practice Through Journal Writing: Impacts of variations in the focus and level of feedback. *Teachers and Teaching*, 8(2):171–196.

Baker, R. (2010). Data mining for education. *International encyclopedia of education*.

Baker, R. S. and Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In Larusson, J. A. and White, B., editors, *Learning Analytics*, pages 61–75. Springer New York.

Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17.

Ballard, K. K. (2006). *Using Van Manen's model to assess levels of reflectivity among preservice physical education teachers*. PhD thesis, Texas A&M University.

Bates, A. W. (2015). *Teaching in a digital age*. open.bccampus.ca.

Batista, G. E., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1):20–29.

Bell, A., Kelton, J., McDonagh, N., Mladenovic, R., and Morrison, K. (2011). A critical evaluation of the usefulness of a coding scheme to categorise levels of reflective thinking. *Assessment & Evaluation in Higher Education*, 36(7):797–815.

Benoit, K., Conway, D., Laver, M., and Mikhaylov, S. (2012). Crowd-sourced data coding for the social sciences: massive non-expert coding of political texts. Havard.

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. (2010). Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, UIST '10, pages 313–322, New York, NY, USA. ACM.

Bienkowski, M., Feng, M., and Means, B. (2012). Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief. Technical report, U.S. Department of Education, Office of Educational Technology, Washington, D.C.

Birney, R. (2012). *Reflective Writing: Quantitative Assessment and Identification of Linguistic Features*. PhD thesis, Waterford Institute of Technology.

Blake, C. (2011). Text mining. *Annual Review of Information Science and Technology*, 45(1):121–155.

Bloom, B. S. (1954). *Taxonomy of educational objectives*. Longmans, Green.

Boenink, A. D., Oderwald, A. K., De Jonge, P., Van Tilburg, W., and Smal, J. A. (2004). Assessing student reflection in medical practice. The development of an observer-rated instrument: reliability, validity and initial experiences. *Medical Education*, 38(4):368–377.

Bogo, M., Regehr, C., Katz, E., Logie, C., and Mylopoulos, M. (2011). Developing a Tool for Assessing Students' Reflections on Their Practice. *Social Work Education*, 30:186–194.

Bond, J. (2003). The Effects of Reflective Assessment on Student Achievement. *Theses and Dissertations*.

Boud, D. (1994). Conceptualising learning from experience: Developing a model for facilitation. In *Proceedings of the 35th Adult Education Research Conference*, pages 49–54, Knoxville, Tennessee.

Boud, D., Keogh, R., and Walker, D. (1985). *Reflection: Turning Experience into Learning*. Routledge.

Boyd, E. M. and Fales, A. W. (1983). Reflective Learning. *Journal of Humanistic Psychology*, 23(2):99 –117.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Brank, J., Mladenic, D., and Grobelnik, M. (2011). Feature Construction in Text Mining. In Sammut, C. and Webb, G. I., editors, *Encyclopedia of Machine Learning*, pages 397–401. Springer US, Boston, MA.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC, New York, N.Y., new ed edition edition.

Brown, G. W. and Hayden, G. F. (1985). Nonparametric Methods: Clinical Applications. *Clinical Pediatrics*, 24(9):490–498.

Bruno, A., Galuppo, L., and Gilardi, S. (2011). Evaluating the reflexive practices in a learning experience. *European Journal of Psychology of Education*, 26:527–543.

Caruana, R. and Niculescu-Mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 161–168, New York, NY, USA. ACM.

Chamoso, J. M. and Cáceres, M. J. (2009). Analysis of the reflections of student-teachers of mathematics when working with learning portfolios in Spanish university classrooms. *Teaching and Teacher Education*, 25(1):198–206.

Chang, C. and Chou, P. (2011). Effects of reflection category and reflection quality on learning outcomes during web-based portfolio assessment process: A case study of high school students in computer application courses. *TOJET*, 10(3).

Chaumba, J. (2015). Using Blogs to Stimulate Reflective Thinking in a Human Behavior Course. *Social Work Education*, 0(0):1–14.

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 853–867. Springer.

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6.

Chirema, K. D. (2007). The use of reflective journals in the promotion of reflection and learning in post-registration nursing students. *Nurse Education Today*, 27(3):192–202.

Chretien, K., Goldman, E., and Faselis, C. (2008). The Reflective Writing Class Blog: Using Technology to Promote Reflection and Professional Development. *Journal of General Internal Medicine*, 23(12):2066–2070.

Chrzaszcz, A., Sporer, T., Metscher, J., Wild, F., and Sigurdarson, S. E. (2008). Distributed e-portfolios to recognise informal learning. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, volume 2008, pages 5830–5838.

Chung, C. K. and Pennebaker, J. W. (2012). Linguistic Inquiry and Word Count (LIWC): Pronounced "Luke" and Other Useful Facts. In McCarthy, P. M. and Boonthum, C., editors, *Applied Natural Language Processing and Content Analysis: Advances in Identification, Investigation and Resolution*, pages 206–229. Information Science Reference (an imprint of IGI Global), Hershey, United States of America.

Clarkeburn, H. and Kettula, K. (2011). Fairness and using reflective journals in assessment. *Teaching in Higher Education*, 17(4):439–452.

Clow, D. (2013). An overview of learning analytics. *Teaching in Higher Education*, 18(6):683–695.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Cohen, W. W. (2004). MinorThird: Methods for Identifying Names and Ontological Relations in Text using Heuristics for Inducing Regularities from Data.

Cohen-Sayag, E. and Fischl, D. (2012). Reflective Writing in Pre-Service Teachers' Teaching: What Does It Promote? *Australian Journal of Teacher Education*, 37(10).

Corich, S., Kinshuk, and Hunt, M., L. (2006). Measuring Critical Thinking within Discussion Forums using a Computerised Content Analysis Tool. *Proceedings of Networked Learning*.

Corich, S. P. (2011). *Automating the measurement of critical thinking in discussion forums*. PhD thesis.

Corlett, S. (2013). Participant learning in and through research as reflexive dialogue: Being 'struck' and the effects of recall. *Management Learning*, 44(5):453–469.

Crawford, S., O'Reilly, R., and Luttrell, S. (2012). Assessing the effects of integrating the reflective framework for teaching in physical education (RFTPE) on the teaching and learning of undergraduate sport studies and physical education students. *Reflective Practice*, 13(1):115–129.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Cronbach, L. J. and Shavelson, R. J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, 64(3):391 –418.

Dascalu, M. (2014). *Analyzing Discourse and Text Complexity for Learning and Collaborating*, volume 534 of *Studies in Computational Intelligence*. Springer International Publishing, Cham.

De Liddo, A., Sándor, A., and Buckingham Shum, S. (2012). Contested Collective Intelligence: Rationale, Technologies, and a Human-Machine Annotation Study. *Computer Supported Cooperative Work (CSCW)*, 21(4-5):417–448.

Dessus, P., Trausan-Matu, S., Van Rosmalen, P., and Wild, F. (2009). AIED 2009 Workshops Proceeedings Volume 10: Natural Language Processing in Support of Learning: Metrics, Feedback and Connectivity.

Dewey, J. (1910). *How we think*. Courier Dover Publications (republication in 1997 of the work orginally published in 1910 by D. C. Heath & Co.).

Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. DC Heath Boston.

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Dönmez, P., Rosé, C., Stegmann, K., Weinberger, A., and Fischer, F. (2005). Supporting CSCL with automatic corpus analysis technology. In *Proceedings of the 2005 conference on Computer support for collaborative learning: learning 2005: the next 10 years!*, pages 125–134.

Duke, S. and Appleton, J. (2000). The use of reflection in a palliative care programme: a quantitative study of the development of reflective skills over an academic year. *Journal of Advanced Nursing*, 32(6):1557–1568.

Dyment, J. E. and O'Connell, T. S. (2010). The Quality of Reflection in Student Journals: A Review of Limiting and Enabling Factors. *Innovative Higher Education*, 35:233–244.

Dyment, J. E. and O'Connell, T. S. (2011). Assessing the quality of reflection in student journals: a review of the research. *Teaching in Higher Education*, 16:81–97.

Eraut, M. (1995). Schon Shock: a case for refraining reflection-in-action? *Teachers and Teaching: theory and practice*, 1(1):9–22.

Erkens, G. and Janssen, J. (2008). Automatic coding of dialogue acts in collaboration protocols. *International Journal of Computer-Supported Collaborative Learning*, 3(4):447–470.

Ertmer, P. A. and Newby, T. J. (2013). Behaviorism, cognitivism, constructivism: Comparing critical features from an instructional design perspective. *Performance Improvement Quarterly*, 26(2):43–71.

Etscheidt, S., Curran, C. M., and Sawyer, C. M. (2012). Promoting Reflection in Teacher Preparation Programs: A Multilevel Model. *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, 35(1):7–26.

Feinerer, I. and Hornik, K. (2014). *tm: Text Mining Package*. R package version 0.6.

Feinerer, I., Hornik, K., and Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, 25(5):1–54.

Fenwick, T. J. (2001). Experiential Learning: A Theoretical Critique from Five Perspectives. Information Series No. 385.

Ferguson, R. and Sharples, M. (2014). Innovative Pedagogy at Massive Scale: Teaching and Learning in MOOCs. In Rensing, C., Freitas, S. d., Ley, T., and Munoz-Merino, P. J., editors, *Open Learning and Teaching in Educational Communities*, number 8719 in Lecture Notes in Computer Science, pages 98–111. Springer International Publishing.

Ferguson, R. and Shum, S. (2011). Learning analytics to identify exploratory dialogue within synchronous text chat. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, pages 99–103.

Ferguson, R. and Shum, S. B. (2012). Social learning analytics: five approaches. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 23–33, New York, NY, USA. ACM.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 15:3133–3181.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J., Nyberg, E., and Prager, J. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79.

Findlay, N., Dempsey, S., and Warren-Forward, H. (2010). Validation and use of the Newcastle Reflective Analysis Tool: a three-year longitudinal study of RT students' reflective journals. *Reflective Practice*, 11(1):83–94.

Findlay, N., Dempsey, S. E., and Warren-Forward, H. M. (2009). Development of the Newcastle Reflective Analysis Tool | NOVA. The University of Newcastle's Digital Repository, Development of the Newcastle Reflective Analysis Tool.

Findlay, N., Dempsey, S. E., and Warren-Forward, H. M. (2011). Development and validation of reflective inventories: assisting radiation therapists with reflective practice. *Journal of Radiotherapy in Practice*, 10(01):3–12.

Fischer, F., Wild, F., Sutherland, R., and Zirn, L., editors (2014). *Grand Challenges in Technology Enhanced Learning - Outcomes of the 3rd Alpine Rendez-Vous*. SpringerBriefs in Education. Springer International Publishing.

Fischer, M. A., Haley, H.-L., Saarinen, C. L., and Chretien, K. C. (2011). Comparison of blogged and written reflections in two medicine clerkships. *Medical Education*, 45(2):166–175.

Fleiss, J. L., Levin, B., and Paik, M. C. (2003). The Measurement of Interrater Agreement. In *Statistical Methods for Rates and Proportions*, pages 598–626. John Wiley & Sons, Inc.

Forbes, A. (2011). Evidence of learning in reflective practice: a case study of computer-assisted analysis of students' reflective blogs. *New Zealand Association for Cooperative Education 2011 Conference Proceedings*, pages 11–14.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *J. Mach. Learn. Res.*, 3:1289–1305.

Frank, E. and Witten, I. H. (1998). Generating accurate rule sets without global optimization. Working Paper, University of Waikato, Department of Computer Science.

Friedman, A. and Schoen, L. (2009). Reflective Practice Interventions: Raising Levels of Reflective Judgment. *Action in Teacher Education*, 31(2):61–73.

Fund, Z., Court, D., and Kramarski, B. (2002). Construction and Application of an Evaluative Tool to Assess Reflection in Teacher-Training Courses. *Assessment & Evaluation in Higher Education*, 27(6):485–499.

Gamer, M., Lemon, J., and <puspendra.pusp22@gmail.com>, I. F. P. S. (2012). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.

Gao, C. and Zhou, D. (2013). Minimax Optimal Convergence Rates for Estimating Ground Truth from Crowdsourced Labels. arXiv e-print 1310.5764.

Gardner, S. and Nesi, H. (2013). A Classification of Genre Families in University Student Writing. *Applied Linguistics*, 34(1):25–52.

Garrison, D. R., Anderson, T., and Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education*, 15(1):7–23.

Garrison, J. (1995). Deweyan Pragmatism and the Epistemology of Contemporary Social Constructivism. *American Educational Research Journal*, 32(4):716–740.

Gore, J. M. and Zeichner, K. M. (1991). Action research and reflective teaching in preservice teacher education: A case study from the United States. *Teaching and Teacher Education*, 7(2):119–136.

Graesser, A. C., D'Mello, S., Hu, X., Cai, Z., Olney, A., and Morgan, B. (2012). AutoTutor. In *Applied Natural Language Processing: Identification, Investigation and Resolution*, pages 169–187. IGI Global.

Granberg, C. (2010). Social software for reflective dialogue: questions about reflection and dialogue in student teachers' blogs. *Technology, Pedagogy and Education*, 19(3):345–360.

Greenwood, J. (1993). Reflective practice: a critique of the work of Argyris and Schön. *Journal of Advanced Nursing*, 18(8):1183–1187.

Gulwadi, G. B. (2009). Using reflective journals in a sustainable design studio. *International Journal of Sustainability in Higher Education*, 10(2):96–106.

Gupta, V. and Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, 1(1):60–76.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Gwet, K. L. (2012). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters, 3rd Edition*. Advanced Analytics, LLC, Gaithersburg, MD, 3rd edition edition.

Hahnemann, B. K. (1986). Journal writing: a key to promoting critical thinking in nursing students. *The Journal of Nursing Education*, 25(5):213–215.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18.

Hamann, J. M. (2002). Reflective Practices and Confluent Educational Perspectives: Three Exploratory Studies.

Hatton, N. and Smith, D. (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education*, 11(1):33–49.

Hawkes, M. (2001). An analysis of critically reflective teacher dialogue in asynchronous computer-mediated communication. In *IEEE International Conference on Advanced Learning Technologies, 2001. Proceedings*, pages 247 –250.

Hawkes, M. (2006). Linguistic Discourse Variables as Indicators of Reflective Online Interaction. *American Journal of Distance Education*, 20(4):231–244.

Hawkes, M. and Romiszowski, A. (2001). Examining the reflective outcomes of asynchronous computer-mediated communication on inservice teacher development. *Journal of Technology and Teacher Education*, 9:285–308.

Hearst, M. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and their Applications*, 15(5):22 –37.

Herring, S., Scheidt, L., Bonus, S., and Wright, E. (2004). Bridging the gap: a genre analysis of Weblogs. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences, 2004*, pages 11 pp.–.

Heuboeck, A., Holmes, J., and Nesi, H. (2007). *The BAWE corpus manual*. Technical report, Universities of Warwick, Coventry and Reading.

Hill, H. R. M., Crowe, T. P., and Gonsalvez, C. J. (2015). Reflective dialogue in clinical supervision: A pilot study involving collaborative review of supervision videos. *Psychotherapy Research*, 0(0):1–16.

Hornik, K., Buchta, C., and Zeileis, A. (2009). Open-Source Machine Learning: R Meets Weka. *Computational Statistics*, 24(2):225–232.

Hornik, K., Meyer, D., and Karatzoglou, A. (2006). Support vector machines in R. *Journal of Statistical Software*, 15(9):1–28.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. (2013). Learning Whom to trust with MACE. In *Proceedings of NAACL-HLT*, pages 1120–1130.

Hsieh, H.-F. and Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, 15(9):1277–1288.

Hutchinson, C. J. and Allen, K. W. (1997). The reflection integration model: A process for facilitating reflective learning. *The Teacher Educator*, 32(4):226–234.

Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing*. Continuum International Publishing Group.

Ip, W. Y., Lui, M. H., Chien, W. T., Lee, I. F., Lam, L. W., and Lee, D. (2012). Promoting self-reflection in clinical practice among Chinese nursing undergraduates in Hong Kong. *Contemporary Nurse*, 41(2):253–262.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429.

Jay, J. K. and Johnson, K. L. (2002). Capturing complexity: a typology of reflective practice for teacher education. *Teaching and Teacher Education*, 18(1):73–85.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In Nédellec, C. and Rouveirol, C., editors, *Machine Learning: ECML-98*, number 1398 in Lecture Notes in Computer Science, pages 137–142. Springer Berlin Heidelberg.

Jordan, S. (2014). *E-assessment for learning? Exploring the potential of computer-marked assessment and computer-generated feedback, from short-answer questions to assessment analytics.* PhD thesis, The Open University.

Kalz, M., Ras, E., Junqueira Barbosa, S. D., Chen, P., Cuzzocrea, A., Du, X., Filipe, J., Kara, O., Kotenko, I., Sivalingam, K. M., Slezak, D., Washio, T., and Yang, X., editors (2014). *Computer Assisted Assessment. Research into E-Assessment*, volume 439 of *Communications in Computer and Information Science*. Springer International Publishing, Cham.

Kang, M., Chaudhuri, S., Kumar, R., Wang, Y.-C., Rosé, E. R., Rosé, C. P., and Cui, Y. (2008). Supporting the Guide on the SIDE. In *Proceedings of the 9th international conference on Intelligent Tutoring Systems*, ITS '08, pages 793–795, Berlin, Heidelberg. Springer-Verlag.

Kansanaho, H., Cordina, M., Puumalainen, I., and Airaksinen, M. (2005). Practicing pharmacists' patient counseling skills in the context of reflectivity. *Pharmacy Education*, 5(1):19–26.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software*, 11(9):1–20.

Kassarjian, H. H. (1977). Content Analysis in Consumer Research. *Journal of Consumer Research*, 4(1):8–18.

Katz, S., O'Donnell, G., and Kay, H. (2000). An approach to analyzing the role and structure of reflective dialogue. *International Journal of Artificial Intelligence in Education (IJAIED)*, 11:320–343.

Kember, D. (2008). *Reflective Teaching and Learning in the Health Professions: Action Research in Professional Education*. John Wiley and Sons.

Kember, D., Jones, A., Loke, A., McKay, J., Sinclair, K., Tse, H., Webb, C., Wong, F., Wong, M., and Yeung, E. (1999). Determining the level of reflective thinking from students' written journals using a coding scheme based on the work of Mezirow. *International Journal of Lifelong Education*, 18(1):18–30.

Kember, D., Leung, D. Y. P., Jones, A., Loke, A. Y., McKay, J., Sinclair, K., Tse, H., Webb, C., Wong, F. K. Y., Wong, M., and Yeung, E. (2000). Development of a Questionnaire to Measure the Level of Reflective Thinking. *Assessment & Evaluation in Higher Education*, 25(4):381.

Kember, D., McKay, J., Sinclair, K., and Wong, F. K. Y. (2008). A four-category scheme for coding and assessing the level of reflection in written work. *Assessment & Evaluation in Higher Education*, 33:369–379.

Kent, A. and McCarthy, P. M. (2012). Discourse Analysis and ANLP. In McCarthy, P. M. and Boonthum-Denecke, C., editors, *Applied Natural Language Processing and Content Analysis: Advances in Identification, Investigation and Resolution*, pages 33–52.

Khan, A., Baharudin, B., Lee, L. H., and Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, 1(1).

Killion, J. P. and Todnem, G. R. (1991). A Process for Personal Theory Building. *Educational Leadership*, 48(6):14–16.

Kim, H. S. (1999). Critical reflective inquiry for knowledge development in nursing practice. *Journal of Advanced Nursing*, 29(5):1205–1212.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11):3735–3745.

King, P. M. and Kitchener, K. S. (1994). *Developing Reflective Judgment*. Jossey-Bass Publishers, San Francisco.

Kittur, A., Chi, E. H., and Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA. ACM.

Kittur, A., Nickerson, J. V., Bernstein, M. S., Gerber, E. M., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. J. (2012). The Future of Crowd Work. SSRN Scholarly Paper ID 2190946, Social Science Research Network, Rochester, NY.

Korthagen, F. and Vasalos, A. (2005). Levels in reflection: core reflection as a means to enhance professional growth. *Teachers and Teaching: Theory and Practice*, 11:47–71.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica 31:249–268. *Informatica*, 31:249–268.

Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2007). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190.

Kovanovic, V., Joksimovic, S., Gasevic, D., and Hatala, M. (2014). Automated cognitive presence detection in online discussion transcripts. In *Proceedings of the Workshops at the LAK 2014 Conference*, volume 1137, Indianapolis, Indiana, USA. CEUR-WS.org.

Kreber, C. (2005). Reflection on teaching and the scholarship of teaching: Focus on science instructors. *Higher Education*, 50:323–359.

Kreber, C. and Castleden, H. (2008). Reflection on teaching and epistemological structure: reflective and critically reflective processes in 'pure/soft' and 'pure/hard' fields. *Higher Education*, 57(4):509–531.

Krippendorff, K. (2004a). Measuring the Reliability of Qualitative Text Analysis Data. *Quality & Quantity*, 38(6):787–800.

Krippendorff, K. (2004b). Reliability in Content Analysis. *Human Communication Research*, 30(3):411–433.

Krippendorff, K. H. (2012). *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Inc, third edition edition.

Kuhn, M., Weston, S., Coulter, N., and Culp, M. (2014a). *C50: C5.0 Decision Trees and Rule-Based Models*. R package version 0.1.0-19.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., and the R. Core Team (2014b). *caret: Classification and Regression Training*. R package version 6.0-30.

Lai, G. and Calandra, B. (2010). Examining the effects of computer-based scaffolds on novice teachers' reflective journal writing. *Etr&d-Educational Technology Research and Development*, 58(4):421–437.

Lambe, J. (2011). Developing pre-service teachers' reflective capacity through engagement with classroom-based research. *Reflective Practice*, 12(1):87–100.

Landauer, T. K. (2003). Automatic Essay Assessment. *Assessment in Education: Principles, Policy & Practice*, 10(3):295–308.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Le Cornu, A. (2009). Meaning, Internalization, and Externalization: Toward a Fuller Understanding of the Process of Reflection and Its Role in the Construction of the Self. *Adult Education Quarterly*, 59(4):279–297.

Lee, H.-J. (2005). Understanding and assessing preservice teachers' reflective thinking. *Teaching and Teacher Education*, 21(6):699–715.

Leijen, Ä., Valtna, K., Leijen, D. A., and Pedaste, M. (2012). How to determine the quality of students' reflections? *Studies in Higher Education*, 37(2):203–217.

Lethbridge, K., Andrusyszyn, M.-A., Iwasiw, C., Laschinger, H. K., and Fernando, R. (2011). Assessing the psychometric properties of Kember and Leung's Reflection Questionnaire. *Assessment & Evaluation in Higher Education*, pages 1–23.

Li, H., Yu, B., and Zhou, D. (2013). Error Rate Bounds in Crowdsourcing Models. In *ICML13 Workshop: Machine Learning Meets Crowdsourcing*.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3):18–22.

Lisacek, F., Chichester, C., Kaplan, A., and Sandor, Á. (2005). Discovering paradigm shift patterns in biomedical abstracts: application to neurodegenerative diseases. In *First international symposium on semantic mining in biomedicine*, pages 11–13.

Lombard, M., Snyder-Duch, J., and Bracken, C. C. (2002). Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, 28(4):587–604.

Luk, J. (2008). Assessing teaching practicum reflections: Distinguishing discourse features of the "high" and "low" grade reports. *System*, 36(4):624–641.

MacLellan, E. (2004). How reflective is the academic essay? *Studies in Higher Education*, 29(1):75–89.

Mamede, S. and Schmidt, H. G. (2004). The structure of reflective practice in medicine. *Medical Education*, 38(12):1302–1308.

Manen, M. v. (1977). Linking Ways of Knowing with Ways of Being Practical. *Curriculum Inquiry*, 6(3):205–228.

Mann, K., Gordon, J., and MacLeod, A. (2007). Reflection and reflective practice in health professions education: a systematic review. *Advances in Health Sciences Education*, 14:595–621.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.

Mansvelder-Longayroux, D., Beijaard, D., and Verloop, N. (2007). The portfolio as a tool for stimulating reflection by student teachers. *Teaching and teacher education*, 23(1):47–62.

Mansvelder-Longayroux, D. D. (2006). *The learning portfolio as a tool for stimulating reflection by student teachers*. Doctoral thesis, ICLON, Leiden University Graduate School of Teaching, Leiden University.

Mayfield, E. and Penstein-Rosé, C. (2010). Using Feature Construction to Avoid Large Feature Spaces in Text Classification. In *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation*, GECCO '10, pages 1299–1306, New York, NY, USA. ACM.

McCarthy, P. M. and Boonthum-Denecke, C. (2012). *Applied natural language processing: identification, investigation, and resolution*. Information Science Reference, Hershey, PA.

McCollum, S. (1997). Insights Into the Process of Guiding reflection During an Early Field experience of Preservice Teachers.

McDonald, P., Straker, H., Schlumpf, K., and Plack, M. (2014). Learning Partnership: Students and Faculty Learning Together to Facilitate Reflection and Higher Order Thinking in a Blended Course. *Online Learning: Official Journal of the Online Learning Consortium*, 18(4).

McKay, F. H. and Dunn, M. (2015). Student reflections in a first year public health and health promotion unit. *Reflective Practice*, 0(0):1–12.

McKlin, T. (2004). Analyzing cognitive presence in online courses using an artificial neural network. *Middle-Secondary Education and Instructional Technology Dissertations*, page 1.

Means, B., Toyama, Y., Murphy, R., Bakia, M., and Jones, K. (2010). Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies. Monograph.

Medwell, J. and Wray, D. (2014). Pre-service teachers undertaking classroom research: developing reflection and enquiry skills. *Journal of Education for Teaching*, 40(1):65–77.

Mena-Marcos, J., García-Rodríguez, M.-L., and Tillema, H. (2013). Student teacher reflective writing: what does it reveal? *European Journal of Teacher Education*, 36(2):147–163.

Menardi, G. and Torelli, N. (2012). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):92–122.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2014). *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.6-3.

Mezirow, J. (1981). A Critical Theory of Adult Learning and Education. *Adult Education Quarterly*, 32:3–24.

Mezirow, J. (1990a). *Fostering critical reflection in adulthood: a guide to transformative and emancipatory learning*. Jossey-Bass.

Mezirow, J. (1990b). How critical reflection triggers transformative learning. In Mezirow, J., editor, *Fostering critical reflection in adulthood: a guide to transformative and emancipatory learning*, pages 1–20. Jossey-Bass.

Mezirow, J. (1991). *Transformative dimensions of adult learning*. Jossey-Bass, 350 Sansome Street, San Francisco, CA 94104-1310 ($27.95).

Mezirow, J. (1998). On Critical Reflection. *Adult Education Quarterly*, 48(3):185–198.

Miles, M. and Huberman, A. (1984). *Qualitative data analysis: A sourcebook of new methods*. Sage publications.

Minott, M. (2008). Valli's Typology Of Reflection And The Analysis Of Pre-Service Teachers' Reflective Journals. *Australian Journal of Teacher Education*, 33(5).

Mladenic, D. (2011). Feature Selection in Text Mining. In Sammut, C. and Webb, G. I., editors, *Encyclopedia of Machine Learning*, pages 406–410. Springer US.

Modupeoluwa, A. Y. (2011). *Intelligent Blogs for Reflection*. BCS Computer Science Thesis, University of Leeds, Leeds.

Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307.

Moon, J. A. (1999). *Reflection in learning & professional development*. Routledge.

Moon, J. A. (2004). *A handbook of reflective and experiential learning*. Routledge.

Moon, J. A. (2006). *Learning Journals: A Handbook for Reflective Practice and Professional Development*. Routledge, 2 edition.

Moseley, D., Baumfield, V., Elliott, J., Gregson, M., Higgins, S., Miller, J., and Newton, D. P. (2005a). *Frameworks for Thinking: A Handbook for Teaching and Learning*. Cambridge University Press.

Moseley, D., Baumfield, V., Higgins, S., Lin, M., Miller, J., Newton, D., Robson, S., Elliott, J., and Gregson, M. (2004). *Thinking Skill Frameworks for Post-16 Learners: An Evaluation. A Research Report for the Learning and Skills Research Centre.* Learning and Skills Development Agency.

Moseley, D., Elliott, J., Gregson, M., and Higgins, S. (2005b). Thinking Skills Frameworks for Use in Education and Training. *British Educational Research Journal*, 31(3):367–390.

Müller, R. and Büttner, P. (2006). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine*, 13(23-24):2465–2476.

Nesi, H. and Edwardes, M. (2007). The form, meaning and purpose of university level assessed reflective writing. In *Proceedings of the BAAL Annual Conference*.

Nesi, H. and Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge University Press.

Nguyen, Q. D., Fernandez, N., Karsenti, T., and Charlin, B. (2014). What is reflection? A conceptual analysis of major definitions and a proposal of a five-component model. *Medical Education*, 48(12):1176–1189.

Noss, R., Cox, R., Laurillard, D., Luckin, R., Plowman, L., Scanlon, E., and Sharples, M. (2012). System upgrade: realising the vision for UK education. Technical report.

O'Connell, T. S. and Dyment, J. E. (2004). Journals of post secondary outdoor recreation students: The results of a content analysis. *Journal of Adventure Education & Outdoor Learning*, 4(2):159–171.

OECD (2013). *PISA 2012 Assessment and Analytical Framework*. PISA. OECD Publishing.

Page, E. B. (1968). The Use of the Computer in Analyzing Student Essays. *International Review of Education / Internationale Zeitschrift für Erziehungswissenschaft / Revue Internationale de l'Education*, 14(2):210–225.

Page, E. B. and Paulus, D. H. (1968). The Analysis of Essays by Computer. Final Report.

Papamitsiou, Z. and Economides, A. A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Journal of Educational Technology & Society*, 17(4).

Paterson, B. L. (1995). Developing and maintaining reflection in clinical journals. *Nurse Education Today*, 15(3):211–220.

Pee, B., Woodman, T., Fry, H., and Davenport, E. S. (2002). Appraising and assessing reflection in students' writing on a structured worksheet. *Medical Education*, 36(6):575–585.

Peltier, J. W., Hay, A., and Drago, W. (2005). The Reflective Learning Continuum: Reflecting on Reflection. *Journal of Marketing Education*, 27(3):250–263.

Pennebaker, J. W. and Francis, M. E. (1996). Cognitive, Emotional, and Language Processes in Disclosure. *Cognition & Emotion*, 10(6):601–626.

Perkins, C. and Murphy, E. (2006). Identifying and measuring individual engagement in critical thinking in online discussions: An exploratory case study. *Journal of Educational Technology And Society*, 9(1):298.

Plack, M., Driscoll, M., Blissett, S., McKenna, R., and Plack, T. (2005). A method for assessing reflective journal writing. *Journal of allied health*, 34(4):199–208.

Plack, M. M., Driscoll, M., Marquez, M., Cuppernull, L., Maring, J., and Greenberg, L. (2007). Assessing Reflective Writing on a Pediatric Clerkship by Using a Modified Bloom's Taxonomy. *Ambulatory Pediatrics*, 7(4):285–291.

Poldner, E., Simons, P., Wijngaards, G., and van der Schaaf, M. (2012). Quantitative content analysis procedures to analyse students' reflective essays: A methodological review of psychometric and edumetric aspects. *Educational Research Review*, 7(1):19–37.

Poldner, E., Van der Schaaf, M., Simons, P. R.-J., Van Tartwijk, J., and Wijngaards, G. (2014). Assessing student teachers' reflective writing through quantitative content analysis. *European Journal of Teacher Education*, 37(3):348–373.

Poom-Valickis, K. and Mathews, S. (2013). Reflecting others and own practice: an analysis of novice teachers' reflection skills. *Reflective Practice*, 14(3):420–434.

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC., Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Prilla, M. and Renner, B. (2014). Supporting Collaborative Reflection at Work: A Comparative Case Analysis. In *Proceedings of the 18th International Conference on Supporting Group Work*, GROUP '14, pages 182–193, New York, NY, USA. ACM Press.

Pultorak, E. G. (1996). Followling the Developmental Process of Reflection in Novice Teachers: Three Years of Investigation. *Journal of Teacher Education*, 47(4):283–291.

QAA (2012). UK quality code for higher education. Part B: Assuring and enhancing academic quality. Chapter B3: Learning and teaching. Technical report.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Quinn, A. J. and Bederson, B. B. (2009). A taxonomy of distributed human computation. *Human-Computer Interaction Lab Tech Report, University of Maryland*.

Quinn, A. J. and Bederson, B. B. (2010). Human computation: Charting the growth of a burgeoning field. *computer*, 1(3):10–37.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ratkic, A. (2012). Images of reflection: on the meanings of the word reflection in different learning contexts. *AI & SOCIETY*.

Reidsema, C. and Mort, P. (2009). Assessing reflective writing: Analysis of reflective writing in an engineering design course. *Journal of Academic Language and Learning*, 3(2):A117–A129.

Revelle, W. (2013). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.3.10.

Richardson, G. and Maltby, H. (1995). Reflection-on-practice: enhancing student learning. *Journal of Advanced Nursing*, 22(2):235–242.

Riffe, D., Lacy, S., and Fico, F. G. (2005). *Analyzing Media Messages: Using Quantitative Content Analysis In Research*. Psychology Press.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77.

Rogers, R. R. (2001). Reflection in Higher Education: A Concept Analysis. *Innovative Higher Education*, 26(1):37–57.

Römer, U. and O'Donnell, M. B. (2011). From student hard drive to web corpus (part 1): the design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 6(2):159–177.

Romero, C. and Ventura, S. (2006). *Data mining in e-learning*. WIT.

Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1):135–146.

Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., and Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, 3(3):237–271.

Ross, D. D. (1989). First Steps in Developing A Reflective Approach. *Journal of Teacher Education*, 40(2):22–30.

Rourke, L., Anderson, T., Garrison, D. R., and Archer, W. (2001). Methodological issues in the content analysis of computer conference transcripts. *International Journal of Artificial Intelligence in Education (IJAIED)*, 12:8–22.

Ryan, M. (2011). Improving reflective writing in higher education: a social semiotic perspective. *Teaching in Higher Education*, 16(1):99–111.

Ryan, M. (2012). Conceptualising and teaching discursive and performative reflection in higher education. *Studies in Continuing Education*, 34(2):207–223.

Ryan, M. (2014). Reflexive writers: Re-thinking writing development and assessment in schools. *Assessing Writing*, 22:60–74.

Rychen, D. and Salganik, L. (2005). *The Definition and Selection of Key Competencies: Executive Summary*. OECD.

Sandor, A. (2005). A framework for detecting contextual concepts in texts. In *ELECTRA Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications (Beyond Bag of Words)*, pages 80–82, Pestana Bahia, Salvador, Brazil.

Sándor, Á. (2006). Using the author's comments for knowledge discovery. *Semaine de la connaissance, Atelier texte et connaissance, Nantes*.

Sandor, A. (2007). Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. *http://www.cairn.info/*, Vol. XII(2):97–108.

Sándor, Á. and Vorndran, A. (2009). Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, pages 36–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Scanlan, J. M. and Chernomas, W. M. (1997). Developing the reflective teacher. *Journal of Advanced Nursing*, 25(6):1138–1143.

Schnoebelen, T. and Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, 43(4):441–464.

Schön, D. (1987). *Educating the reflective practitioner*. Jossey-Bass San Francisco.

Schön, D. A. (1983). *The reflective practitioner*. Basic Books New York.

Schuh, K. L. and Barab, S. A. (2008). Philosophical Perspectives. In Spector, J. M., Merrill, M. D., Merrienboer, J. v., and Driscoll, M. P., editors, *Handbook of Research on Educational Communications and Technology*, pages 67–82. Routledge, New York, 3 edition.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Comput. Surv.*, 34(1):1–47.

Shaheed, N. and Dong, A. (2006). Reflection and analysis in design student blogs. *DS 36: Proceedings DESIGN 2006, the 9th International Design Conference, Dubrovnik, Croatia*.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20:53–76.

Shermis, M. D. and Burstein, J. C., editors (2003). *Automated Essay Scoring: A Cross-disciplinary Perspective*. Routledge.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Siegel, S. (1957). Nonparametric Statistics. *The American Statistician*, 11(3):13–19.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263.

Sobral, D. T. (2000). An appraisal of medical students' reflection-in-learning. *Medical Education*, 34(3):182–187.

Sobral, D. T. (2005). Medical Students' Mindset for Reflective Learning: A Revalidation Study of the Reflection-In-Learning Scale. *Advances in Health Sciences Education*, 10(4):303–314.

Spalding, E., Wilson, A., and Mewborn, D. (2002). Demystifying reflection: A study of pedagogical strategies that encourage reflective journal writing. *The Teachers College Record*, 104(7):1393–1421.

Sparks-Langer, G. M. and Colto, A. B. (1991). Synthesis of Research on Teachers' Reflective Thinking. *Educational Leadership*, 48(6):37–44.

Sparks-Langer, G. M., Simmons, J. M., Pasch, M., Colton, A., and Starko, A. (1990). Reflective Pedagogical Thinking: How Can We Promote It and Measure It? *Journal of Teacher Education*, 41(5):23 –32.

Stemler, S. E. and Tsai, J. (2008). Best Practices in Interrater Reliability Three Common Approaches. In *Best Practices in Quantitative Methods*, pages 29–49. SAGE Publications, Inc., 2455 Teller Road, Thousand Oaks California 91320 United States of America.

Stone, P. J. and Hunt, E. B. (1963). A computer approach to content analysis: studies using the General Inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA. ACM.

Sumsion, J. and Fleet, A. (1996). Reflection: can we assess it? Should we assess it? *Assessment & Evaluation in Higher Education*, 21(2):121.

Surbeck, E., Han, E. P., and Moyer, J. E. (1991). Assessing Reflective Responses in Journals. *Educational Leadership*, 48(6):25–27.

Sutherland, R., Eagle, S., and Joubert, M. (2012). A vision and strategy for Technology Enhanced Learning: Report from the STELLAR Network of Excellence. Technical report.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Taylor, E. W. (1997). Building Upon the Theoretical Debate: A Critical Review of the Empirical Studies of Mezirow's Transformative Learning Theory. *Adult Education Quarterly*, 48:34–59.

Teufel, S. (1999). *Argumentative zoning: Information extraction from scientific text*. PhD thesis.

Therneau, T., Atkinson, B., and Ripley, B. (2014). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-8.

Thorpe, K. (2004). Reflective learning journals: From concept to practice. *Reflective Practice*, 5(3):327–343.

Tillema, H. H. (2004). The Dilemma of Teacher Educators: Building actual teaching on conceptions of learning to teach. *Teaching Education*, 15(3):277–291.

Tsangaridou, N. and O'Sullivan, M. (1994). Using pedagogical reflective strategies to enhance reflection among pre service physical education teachers.

Ullmann, T. D. (2011). An Architecture for the Automated Detection of Textual Indicators of Reflection. In Reinhardt, W., Ullmann, T. D., Scott, P., Pammer, V., Conlan, O., and Berlanga, A., editors, *Proceedings of the 1st European Workshop on Awareness and Reflection in Learning Networks*, pages 138–151, Palermo, Italy. CEUR-WS.org.

Ullmann, T. D., Wild, F., and Scott, P. (2012). Comparing Automatically Detected Reflective Texts with Human Judgements. In Moore, A., Pammer, V., Pannese, L., Prilla, M., Rajagopal, K., Reinhardt, W., Ullmann, T. D., and Voigt, C., editors, *2nd Workshop on Awareness and Reflection in Technology-Enhanced Learning*, Saarbruecken, Germany. CEUR-WS.org.

Ullmann, T. D., Wild, F., and Scott, P. (2013). Reflection - quantifying a rare good. In Kravcik, M., Krogstie, B. R., Moore, A., Pammer, V., Pannese, L., Prilla, M., Reinhardt, W., and Ullmann, T. D., editors, *Proceedings of the 3rd Workshop on Awareness and Reflection in Technology-Enhanced Learning*, pages 29–40, Paphos, Cyprus. CEUR-WS.org.

Valli, L. (1997). Listening to Other Voices: A Description of Teacher Reflection in the United States. *Peabody Journal of Education*, 72(1):67–88.

Van Manen, M. (1977). Linking Ways of Knowing with Ways of Being Practical. *Curriculum Inquiry*, 6(3):205–228.

Van Manen, M. (1995). On the epistemology of reflective practice. *Teachers and Teaching: theory and practice*, 1(1):33–50.

van Merriënboer, J. J. G. and de Bruin, A. B. H. (2014). Research Paradigms and Perspectives on Learning. In Spector, J. M., Merrill, M. D., Elen, J., and Bishop, M. J., editors, *Handbook of Research on Educational Communications and Technology*, pages 21–29. Springer New York, New York, NY, 4 edition.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Vermunt, J. D. and Verloop, N. (1999). Congruence and friction between learning and teaching. *Learning and Instruction*, 9(3):257–280.

Wald, H. S., Borkan, J. M., Taylor, J. S., Anthony, D., and Reis, S. P. (2012). Fostering and Evaluating Reflective Capacity in Medical Education: Developing the REFLECT Rubric for Assessing Reflective Writing. *Academic Medicine*, 87(1):41–50.

Wallman, A., Lindblad, A. K., Hall, S., Lundmark, A., and Ring, L. (2008). A Categorization Scheme for Assessing Pharmacy Students' Levels of Reflection During Internships. *American Journal of Pharmaceutical Education*, 72(1).

Wang, A., Hoang, C. D. V., and Kan, M.-Y. (2013). Perspectives on crowdsourcing annotations for natural language processing. *Language Resources and Evaluation*, 47(1):9–31.

Ward, J. R. and McCotter, S. S. (2004). Reflection as a visible outcome for preservice teachers. *Teaching and Teacher Education*, 20(3):243–257.

Weihs, C., Ligges, U., Luebke, K., and Raabe, N. (2005). klaR Analyzing German Business Cycles. In Baier, D., Decker, R., and Schmidt-Thieme, L., editors, *Data Analysis and Decision Support*, pages 335–343, Berlin. Springer-Verlag.

Weinberger, A. and Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46(1):71–95.

Wessel, J. and Larin, H. (2006). Change in reflections of physiotherapy students over time in clinical placements. *Learning in Health and Social Care*, 5(3):119–132.

Wharton, S. (2012). Presenting a united front: assessed reflective writing on a group experience. *Reflective Practice*, 13(4):489–501.

White, H. D. (2011). Scientific and scholarly networks. In Scott, J. and Carrington, P. J., editors, *The SAGE Handbook of Social Network Analysis*, pages 271–285.

Wild, F., Stahl, C., Stermsek, G., Penya, Y. K., and Neumann, G. (2005). Factors Influencing Effectiveness in Automated Essay Scoring with LSA. In *AIED*, pages 947–949.

Williams, R. M., Wessel, J., Gemus, M., and Foster-Seargeant, E. (2002). Journal writing to promote reflection by physical therapy students during clinical placements. *Physiotherapy Theory & Practice*, 18(1):5–15.

Wilson, J. (2008). Reflecting-on-the-future: a chronological consideration of reflective practice. *Reflective Practice*, 9:177–184.

Winkler, R. L. and Clemen, R. T. (2004). Multiple Experts vs. Multiple Methods: Combining Correlation Assessments. *Decision Analysis*, 1(3):167–176.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2nd edition.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Woerkom, M. v. and Croon, M. (2008). Operationalising critically reflective work behaviour. *Personnel Review*, 37(3):317–331.

Woerkom, M. v., Nijhof, W. J., and Nieuwenhuis, L. F. (2002). Critical reflective working behaviour: a survey research. *Journal of European Industrial Training*, 26(8):375–383.

Wong, F. K., Kember, D., Chung, L. Y. F., and Yan, L. (1995). Assessing the level of student reflection from reflective journals. *Journal of Advanced Nursing*, 22(1):48–57.

Yuen, M.-C., Chen, L.-J., and King, I. (2009). A Survey of Human Computation Systems. In *International Conference on Computational Science and Engineering, 2009. CSE '09*, volume 4, pages 723–728.

Zeichner, K. and Liston, D. (1987). Teaching student teachers to reflect. *Harvard Educational Review*, 57(1):23–49.